

统计分析系列

SPSS 23 (中文版)

统计分析实用教程

(第 2 版)

邓维斌 周玉敏 编著
刘 进 田帅辉

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书是作者经过 10 余年的教学实践,在总结前期版本教材的基础上,根据广大读者的反馈意见修订而成的。全书基于 SPSS 23 中文版软件版本,同时兼顾早期的软件版本。在编写过程中,作者以统计分析的实际应用为主线,在对主要统计分析方法的基本概念和统计学原理进行简明介绍的基础上,以 64 个实例为载体对 SPSS 23 中各种分析方法的应用场合和操作过程进行了清晰说明,并对相关领域的 29 个统计分析典型案例进行了应用方法及解决思路的详细分析。全书共有思考与练习题 76 个,以供巩固学习效果和课后练习。

全书内容覆盖了 SPSS 23 中常用的统计分析方法,共 13 章。第 1 章介绍 SPSS 的基础知识;第 2 章介绍统计数据的收集与预处理;第 3~12 章介绍 SPSS 23 的各种统计方法,包括描述性统计分析、均值比较与 T 检验、非参数检验、方差分析、相关分析、回归分析、聚类 and 判别分析、主成分分析和因子分析、时间序列分析、信度分析;第 13 章介绍图表的创建与编辑。与教材配套的资源包括所有实例、典型案例和习题的数据文件,课程 PPT 教案,以及思考与练习题的参考答案,可登录华信教育资源网 www.hxedu.com.cn 免费下载。

本书可作为高等院校经济学、管理学、教育学、心理学等相关专业本科生和研究生教材,也可供从事统计分析和决策的各领域工作者学习参考。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

SPSS 23(中文版)统计分析实用教程 / 邓维斌等编著. —2 版. —北京:电子工业出版社,2017.6

(统计分析系列)

ISBN 978-7-121-31400-1

I. ①S… II. ①邓… III. ①统计分析—软件包—高等学校—教材 IV. ①C819

中国版本图书馆 CIP 数据核字(2017)第 084987 号

策划编辑:秦淑灵

责任编辑:秦淑灵

印 刷:

装 订:

出版发行:电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本: 787×1092 1/16 印张: 20.75 字数: 598 千字

版 次: 2012 年 3 月第 1 版

2017 年 6 月第 2 版

印 次: 2017 年 6 月第 1 次印刷

印 数: 3000 册 定价: 45.00 元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010)88254888,88258888。

质量投诉请发邮件至 zlt@phei.com.cn,盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式: qinshl@phei.com.cn。

前 言

SPSS 统计分析软件以其易用性和强大功能已成为目前最流行的统计分析工具之一,在国内具有很大的用户群,是目前国内进行管理决策、市场分析、社会调查、医学统计、金融决策等的统计分析人员应用最广泛的软件。

作者经过 10 余年的教学实践,发现在用 SPSS 软件进行数据分析时存在以下几个突出的问题:

(1) 遇到具体问题时不知道该用何种分析方法,即不知 SPSS 所提供的各种统计分析功能究竟适用于解决何种实际问题;

(2) 不知每一种分析方法的具体操作、分析步骤该如何进行;

(3) 不知如何组织数据,即如何将已有的数据组织成适合于 SPSS 特定分析方法的数据格式,如怎么定义变量,如何进行分组等;

(4) 不知如何对输出结果(包括表和图形)进行分析,对涉及假设检验的问题,分不清原假设和备选假设,不知如何根据所输出的统计量及概率值对其进行假设检验。

针对这些问题,我们在不断总结已有讲义、实验指导书和教材的经验 and 不足的基础上,于 2012 年出版了《SPSS 19(中文版)统计分析实用教程》,深受全国各地老师、学生和数据工作者的广泛好评,居于同类书籍销售排行榜前列。在前期版本的基础上,我们基于 SPSS 23 编写了《SPSS 23(中文版)统计分析实用教程(第 2 版)》,根据软件发展和广大读者的要求,我们对原书作了仔细检查、修正和改写,所作的修订如下:

(1) 本书操作基于 SPSS 23 的软件版本,同时兼顾早期的软件版本。

(2) 对“图表的创建与编辑”一章进行大幅度修订,对常用统计图的绘制进行了较详细的介绍。

(3) 将“参数估计与假设检验”一章拆分成两章,分别是“均值比较与 T 检验”和“非参数检验”,内容介绍更加清晰、明白和有针对性。

(4) 增加了部分章节内容,如“多元方差分析”、“非线性回归分析”、“评分者信度分析”等。

(5) 对部分例题、典型案例、思考与练习题进行了精选,使其更加具有针对性。

该教材吸收了前两版教材的优点,集中体现了如下几个特点:

(1) 基于 SPSS 23 中文版软件,典型案例和习题丰富。

本书以 IBM SPSS Statistics 23 简体中文版为蓝本进行编写,扫除了广大国内读者对英文版教材学习的语言障碍。全书以 SPSS 的实际应用为主线,组织了 64 个实例对各项统计分析方法进行介绍,并对相关领域的 29 个统计分析典型案例进行了应用方法及解决思路等的详细分析,全书共有思考与练习题 76 个,以供巩固学习效果。

(2) 全书结构清晰,体系完整,内容精简明了。

在总体内容把握上,按照“SPSS 23 概述—数据组织—统计分析—图形功能”的顺序组织,由浅入深、由基础到专业。在每章内容的安排上按“分析方法简介—统计原理与步骤—统计实例分析—典型案例—思考与练习”的顺序组织,方便读者学习。全书内容涵盖了 SPSS 中最常用的统计分析方法。同时避免了大而全的介绍,只针对最常用的统计功能进行阐述,使读者在有限的时间内学习到更多的实用功能。

(3) 统计分析方法、SPSS 操作和案例分析的有机结合。

从快速掌握和应用 SPSS 的角度出发,作者将 SPSS 各项功能的操作介绍与统计分析方法论述有机结合。对各种统计分析方法的原理进行了通俗易懂的介绍,但又避免了纷繁复杂的数学证明过程,使读者可以了解分析方法的核心思想,掌握方法的正确应用范围。以实例为载体,比较详细地介绍了 SPSS 中各项常用功能菜单和相应对话框的具体意义和适用情况。最后通过多领域的大量分析案例将 SPSS 的操作和统计分析方法进行有机结合。

(4) 加强对特定问题的分析,以及对数据组织方法和分析结果的讨论。

在用 SPSS 对每个案例进行操作之前,设有“分析”步骤,即针对每个具体问题,对为什么要使用该种分析方法进行了解释和说明,在“数据组织”步骤对特定分析方法如何组织数据做了说明,并对每个案例的主要运行结果进行了详尽的解释和讨论。特别对易混淆的问题以注释的方式进行了说明,以方便读者对相关概念和问题进行区别和理解。

本书可供高等院校相关专业的本科生、研究生作为教材使用,也可作为 SPSS 统计分析培训和自学教材。另外,在统计分析或科研中需处理数据的人士也可以参考。与教材配套的资源有所有实例、典型案例和习题的数据文件,课程 PPT 教案,部分思考与练习题的参考答案,可登录华信教育资源网 <http://www.hxedu.com.cn> 免费注册下载。

本书由重庆邮电大学周玉敏老师执笔编写第 1、2、3、13 章,刘进老师编写第 4、5、7 章,邓维斌老师编写第 6、8、9、10 章,田帅辉老师编写第 11、12 章,由邓维斌负责全书的统稿。在本书的编写过程中,有着丰富教材编写经验的万晓榆、吴先锋、刘跃、卢安文、杜茂康、陈文沛等教授给予了较大的帮助和指导,胡大权、陈家佳等老师做了一些基础工作并提出了建设性的意见和建议,王海锦、齐捷等研究生做了大量校对工作,在此表示衷心的感谢。

该书的出版得到了重庆邮电大学教材建设项目(JC2016-09)、重庆市高等教育教学改革重点项目(132004)、重庆邮电大学校级教改项目(XJG1603)等的资助。此外,在该书的编写过程中借鉴了多种相关书籍,引用了一些宝贵的资料,在此向书籍作者表示深切谢意。

本书仅就 SPSS 23 中常用的统计分析方法进行了介绍,书中所论并不完美,对错误和疏漏之处,恳请读者批评指正。笔者 E-mail: dengwb@cqupt.edu.cn。

编著者
2017 年 1 月

目 录

第 1 章	SPSS 软件概述	1
1.1	SPSS 简介	1
1.1.1	SPSS 的发展	1
1.1.2	SPSS 统计分析软件的特点	1
1.1.3	SPSS 23 的新特性	3
1.1.4	SPSS 的模块	4
1.2	SPSS 使用基础	6
1.2.1	SPSS 的安装	6
1.2.2	SPSS 的界面	7
1.3	SPSS 的帮助系统	10
1.3.1	主题	10
1.3.2	教程	10
1.3.3	个案研究	11
1.3.4	统计辅导	11
1.3.5	高级帮助	12
1.4	利用 SPSS 进行数据分析的步骤	13
1.4.1	统计学中数据分析的一般步骤	13
1.4.2	利用 SPSS 进行数据分析的一般步骤	14
第 2 章	统计数据的收集与预处理	15
2.1	统计数据的收集	15
2.1.1	问卷设计	15
2.1.2	问卷分析	18
2.2	数据文件的建立	18
2.2.1	统计数据的度量尺度	18
2.2.2	SPSS 数据文件的特点	19
2.2.3	输入数据建立数据文件	20
2.2.4	从其他数据文件导入数据建立数据文件	25
2.3	数据文件的编辑	28
2.3.1	数据文件的合并	28
2.3.2	数据文件的拆分	30
2.3.3	数据的选取	34
2.3.4	数据的加权	35
2.4	SPSS 数据加工	37

2.4.1	变量的计算	37
2.4.2	数据可视分箱	38
2.4.3	数据重新编码	41
2.5	思考与练习	42
第 3 章	描述性统计分析	44
3.1	基本描述性统计量简介	44
3.1.1	描述集中趋势的统计量	44
3.1.2	描述离散程度的统计量	45
3.1.3	描述总体分布形态的统计量	46
3.2	频率分析	47
3.2.1	基本概念及统计原理	47
3.2.2	SPSS 实例分析	47
3.3	描述性分析	50
3.3.1	基本概念及统计原理	50
3.3.2	SPSS 实例分析	51
3.4	探索性分析	52
3.4.1	基本概念及统计原理	52
3.4.2	SPSS 实例分析	52
3.5	交叉表分析	57
3.5.1	基本概念及统计原理	57
3.5.2	SPSS 实例分析	58
3.6	多重响应分析	63
3.6.1	基本概念及统计原理	63
3.6.2	多重响应分析 SPSS 实例分析	64
3.7	典型案例	66
3.7.1	城市平均气温基本特征分析	66
3.7.2	商场电视品牌满意度调查	67
3.8	思考与练习	68
第 4 章	均值比较与 T 检验	69
4.1	假设检验	69
4.1.1	基本概念及统计原理	70
4.1.2	小概率事件原理	71
4.1.3	假设检验的一般步骤	71
4.2	平均值分析	72
4.2.1	平均值分析的概念及统计原理	72

4.2.2	平均值 SPSS 实例分析	72
4.3	单样本 T 检验	75
4.3.1	基本概念及统计原理	75
4.3.2	单样本 T 检验 SPSS 实例分析	76
4.4	独立样本 T 检验	77
4.4.1	基本概念及统计原理	77
4.4.2	独立样本 T 检验 SPSS 实例分析	78
4.4.3	摘要独立样本 T 检验	80
4.5	配对样本 T 检验	81
4.5.1	基本概念及统计原理	81
4.5.2	配对样本 T 检验 SPSS 实例分析	82
4.6	典型案例	83
4.6.1	蛋白饲料对小白鼠体重影响分析	83
4.6.2	健康教育对儿童血红蛋白水平的影响分析	84
4.6.3	储户的储蓄金额的差异分析	84
4.7	思考与练习	85
第 5 章	非参数检验	86
5.1	参数检验与非参数检验的比较	86
5.2	单样本的非参数检验	87
5.2.1	基本概念及设置	87
5.2.2	卡方检验	90
5.2.3	二项分布检验	96
5.2.4	游程检验	99
5.2.5	单样本 K-S 检验	103
5.3	独立样本非参数检验	106
5.3.1	基本概念及统计原理	106
5.3.2	独立样本非参数检验 SPSS 实例分析	107
5.4	相关样本的非参数检验	111
5.4.1	基本概念及统计原理	111
5.4.2	相关样本的非参数检验 SPSS 实例分析	112
5.5	典型案例	115
5.5.1	判断某产品的需求量是否服从泊松分布	115
5.5.2	调控政策前后大中城市住宅销售价格指数差异性分析	116

5.5.3	某行业企业赢利比例判断	116
5.5.4	棉条棉结杂质粒数分析	116
5.6	思考与练习	117
第 6 章	方差分析	119
6.1	方差分析简介	119
6.1.1	方差分析的概念	119
6.1.2	方差分析的一般步骤	120
6.2	单因素方差分析	120
6.2.1	基本概念及统计原理	120
6.2.2	单因素方差分析 SPSS 实例分析	121
6.3	多因素方差分析	126
6.3.1	基本概念及统计原理	126
6.3.2	多因素方差分析 SPSS 实例分析	128
6.4	协方差分析	134
6.4.1	基本概念及统计原理	134
6.4.2	协方差分析 SPSS 实例分析	135
6.5	多元方差分析	138
6.5.1	基本概念及统计原理	138
6.5.2	多元方差分析 SPSS 实例分析	138
6.6	典型案例	142
6.6.1	培训材料效果分析	142
6.6.2	火箭射程影响因素分析	142
6.6.3	三种治疗高血压病的方法效果分析	143
6.7	思考与练习	143
第 7 章	相关分析	145
7.1	相关分析简介	145
7.1.1	相关分析的概念	145
7.1.2	相关关系的种类	145
7.2	两变量相关分析	146
7.2.1	基本概念及统计原理	146
7.2.2	两变量相关分析 SPSS 实例分析	148
7.3	偏相关分析	151
7.3.1	基本概念及统计原理	151
7.3.2	偏相关分析 SPSS 实例分析	152
7.4	距离分析	154
7.4.1	基本概念及统计原理	154
7.4.2	距离分析 SPSS 实例分析	154

7.5	典型案例	160	9.4.1	基本概念及统计原理	208
7.5.1	有氧训练中的耗氧量研究	160	9.4.2	系统聚类 SPSS 实例分析	209
7.5.2	控制不良贷款	161	9.5	判别分析	214
7.5.3	学生身体状况指标的相似性 分析	162	9.5.1	基本概念及统计原理	214
7.6	思考与练习	162	9.5.2	判别分析 SPSS 实例分析	215
第 8 章	回归分析	165	9.6	典型案例	220
8.1	回归分析简介	165	9.6.1	美国 22 家企业类型划分	220
8.1.1	回归分析的概念	165	9.6.2	销售地区的选择	221
8.1.2	回归分析的一般步骤	166	9.6.3	地区降水量区域类型判别	222
8.2	线性回归分析	167	9.7	思考与练习	223
8.2.1	基本概念及统计原理	167	第 10 章	主成分分析和因子分析	226
8.2.2	一元线性回归 SPSS 实例分析	168	10.1	主成分分析和因子分析简介	226
8.2.3	多元线性回归 SPSS 实例分析	174	10.1.1	基本概念和主要用途	226
8.3	曲线回归分析	179	10.1.2	主成分和公因子数量的确定	227
8.3.1	基本概念及统计原理	179	10.1.3	主成分分析与因子分析的区别 与联系	228
8.3.2	曲线回归 SPSS 实例分析	180	10.2	主成分分析	228
8.4	非线性回归分析	183	10.2.1	基本概念及统计原理	228
8.4.1	基本概念及统计原理	183	10.2.2	主成分分析 SPSS 实例分析	230
8.4.2	非线性回归 SPSS 实例分析	185	10.3	因子分析	237
8.5	二元 Logistic 回归分析	188	10.3.1	基本概念及统计原理	237
8.5.1	基本概念及统计原理	188	10.3.2	因子分析 SPSS 实例分析	238
8.5.2	二元 Logistic 回归 SPSS 实例 分析	189	10.4	典型案例	241
8.6	典型案例	194	10.4.1	医院工作质量评价分析	241
8.6.1	水稻产量影响因素分析	194	10.4.2	各省、市、自治区城市市政 设施建设状况分析	243
8.6.2	产品废品率的因素拟合	195	10.4.3	大学生的价值观分析	244
8.6.3	高管培训与表现预测	195	10.5	思考与练习	244
8.6.4	肾细胞癌转移的判断	196	第 11 章	时间序列分析	246
8.7	思考与练习	197	11.1	时间序列的建立和平稳化	246
第 9 章	聚类和判别分析	199	11.1.1	填补缺失值	246
9.1	聚类和判别分析简介	199	11.1.2	定义日期变量	247
9.1.1	基本概念	199	11.1.3	创建时间序列	248
9.1.2	样本间亲疏关系的度量	200	11.2	指数平滑法	250
9.2	二阶聚类	201	11.2.1	基本概念及统计原理	250
9.2.1	基本概念及统计原理	201	11.2.2	指数平滑法 SPSS 实例分析	251
9.2.2	二阶聚类 SPSS 实例分析	201	11.3	ARIMA 模型	258
9.3	K-均值聚类	204	11.3.1	基本概念及统计原理	258
9.3.1	基本概念及统计原理	204	11.3.2	ARIMA 实例分析	260
9.3.2	K-均值聚类 SPSS 实例分析	205	11.4	时间序列的季节性分解	269
9.4	系统聚类	208	11.4.1	基本概念及统计原理	269

11.4.2	季节性分解的实例分析	269	13.2.1	图表构建器概述	291
11.5	典型案例	272	13.2.2	使用图表构建器创建图形 举例	291
11.5.1	中国社会消费品零售总额 分析	272	13.3	图形画板模板选择器创建图形	295
11.5.2	中国彩电出口数据分析	273	13.3.1	图形画板模板选择器概述	295
11.5.3	城市温度的季节性分解	273	13.3.2	使用图形画板模板选择器创建 图形举例	295
11.6	思考与练习	274	13.4	使用旧对话框创建图形	297
第 12 章	信度分析	276	13.4.1	条形图	297
12.1	内在信度分析	276	13.4.2	三维条形图	299
12.1.1	基本概念及统计原理	276	13.4.3	折线图	302
12.1.2	内在信度实例分析	277	13.4.4	面积图	305
12.2	再测信度分析	283	13.4.5	饼图	306
12.2.1	基本概念及统计原理	283	13.4.6	盘高-盘低图	308
12.2.2	再测信度实例分析	284	13.4.7	箱图	310
12.3	评分者信度分析	286	13.4.8	误差条图	312
12.3.1	基本概念及统计原理	286	13.4.9	人口金字塔图	314
12.3.2	评分者信度实例分析	287	13.4.10	散点图	316
12.4	典型案例	288	13.4.11	直方图	317
12.4.1	Oxford 学习策略量表信度 分析	288	13.5	图表的编辑	319
12.5	思考与练习	288	13.5.1	图表编辑器布局	319
第 13 章	图表的创建与编辑	289	13.5.2	图表编辑基本方法	320
13.1	SPSS 的图形功能概述	289	13.5.3	图表基本设定	320
13.1.1	SPSS 创建图形的一般过程	289	13.5.4	图表高级设定	321
13.1.2	图形生成与数据文件结构	289	13.6	思考与练习	321
13.1.3	图形生成与数据的度量尺度	290	参考文献		323
13.2	图表构建器创建图形	291			

第 1 章 SPSS 软件概述

在科学研究中，常常需要对大量的数据进行统计处理，这是一项细致而烦琐的工作，如果完全依靠手工来进行，工作量较大，且难以保证准确性，也达不到高精度。为了减轻整理和计算大量数据的负担，提高工作效率，我们必须充分利用现代化的技术手段。随着计算机软件技术的发展，计算机在分析数据方面发挥了相当大的作用，它功能多，速度快，计算精确，较易利用，并且计算机统计软件可以完成更为精确系统的数据分析与统计计算。

在资料统计处理中常采用的统计软件有 SPSS 统计软件系统、SAS 统计分析系统和 Microsoft 公司的 Excel 软件等。SPSS 最初是 Statistics Package for Social Sciences（社会科学统计软件包）的缩写，它是社会科学研究人员首选的统计软件，也是目前世界上最流行的统计软件之一。现随着服务领域的扩大和服务深度的增加，英文已更改为“Statistics Product and Service Solution”，意为“统计产品与服务解决方案”。SPSS 是在经济学、生物学、心理学、医疗卫生、体育、农业、林业、商业、金融等各个领域广泛应用的软件。它不仅能够实现统计功能，还能将分析结果用多种清晰简练的表格和数十种生动形象的二维、三维图形来表达，真正做到实用与美观的统一。

作为全书的开篇，本章介绍 SPSS 的基础知识，主要包括 SPSS 23 的新特性、主要功能模块、常用窗口、帮助系统的使用，以及利用 SPSS 进行数据分析的基本步骤。

1.1 SPSS 简介

1.1.1 SPSS 的发展

SPSS 统计软件系统最早是在 1968 年由美国斯坦福大学的 3 位学生开发的一个软件包，基于这一系统，于 1975 年在芝加哥成立了 SPSS 公司。1984 年，SPSS 首先推出了世界上第一个统计分析软件微机版本 SPSS/PC+，它迅速占领了微机市场，扩大了自己的用户量，开创了 SPSS 微机系列产品的开发方向。

20 世纪 80 年代末，Microsoft 公司发布 Windows 操作系统后，SPSS 迅速向 Windows 移植。至 1993 年 6 月，正式推出 SPSS for Windows 6.0 版本。随后几乎每年推出一个更新版本，2009 年，SPSS 公司将 4 大系列产品（Statistics Family、Modeling Family、Data Collection Family、Deployment Family）整合到一个综合分析平台，把 4 类产品统一加上 PASW（为 Predictive Analysis SoftWare 的首字母）前缀，喻义 SPSS 产品的发展方向为预测分析领域。此后，SPSS 把正在发行的 SPSS 17 统计分析软件正式更名为 PASW Statistics 17，并且，从此开始成为多国语言版本，有了官方的中文界面及使用手册。随后，SPSS 公司被 IBM 收购成其子公司，将 SPSS 统计分析产品更名为 IBM SPSS Statistics。本书以 IBM SPSS Statistics 23 for Windows 为蓝本，结合统计学知识，对各领域常见统计分析案例进行分析讲述。

1.1.2 SPSS 统计分析软件的特点

随着 SPSS 的版本不断更新，软件功能不断完善，操作越来越简便，与其他软件的接口也越来越多。SPSS Statistics 23 for Windows 具有以下特点。

1. Windows 风格，界面友好

SPSS Statistics for Windows 最突出的特点就是操作界面友好，输出结果美观。SPSS 是第一个采用人机交互界面的统计软件，非常容易学习和使用。自从 1995 年 SPSS 公司与微软公司合作开发 SPSS 界面后，SPSS 界面变得越来越友好，操作也越来越简单，这就使熟悉微软公司产品的用户学习 SPSS Statistics 操作时，很容易上手。

SPSS 界面非常友好，熟悉的 Windows 风格界面，数据视图也类似 Excel 布局。具有第四代语言的特点，告诉系统要做什么，无须告诉怎样做。只要了解统计分析的原理，无须通晓统计方法的各种算法，即可得到需要的统计分析结果。除了数据录入及部分命令程序等少数输入工作需要键盘输入外，大多数操作可通过鼠标拖曳、单击“菜单”、“按钮”和“对话框”来完成。

SPSS 功能强大，界面友好，易学易用。SPSS 界面完全是菜单式，使用下拉菜单来选择所需要执行的复杂的统计命令，使用 Windows 的窗口方式展示各种管理和分析数据方法的功能，使用对话框展示出各种功能选择项，只要掌握一定的 Windows 操作技能，粗略了解统计分析原理，就可以使用该软件为特定的科研工作服务。开放式的命令语句窗口，可以通过复制和粘贴的方法来学习和使用其“统计程序”句法语言，同时也适合数据分析专家和研究员使用。

2. 易学易用

SPSS 易于操作，易于入门，结果易于阅读，对统计软件的学习不会冲淡统计的主题，这样研究人员就可以将精力集中在社会研究方法、市场研究方法、营销的业务问题上，而不是忙于编程和统计。一般稍有统计基础的人经过几天的培训即可用 SPSS 做简单的数据分析，包括绘制图表、简单回归、相关分析等。当然，真正应用好 SPSS 软件的关键在于科学地设计研究方案、严谨地收集数据、严密深入地对数据进行统计分析及解释，以及适度保守地下研究结论和进行决策。这一方面要求研究者掌握数理统计的基本知识，另一方面也要求研究者多进行实践，在实践中了解各种统计结果的实际意义。

从某种意义上讲，SPSS 软件还可以帮助数学功底不够扎实的用户学习运用现代统计技术。用户仅需要关心某个问题应该采用何种统计方法，并初步掌握对计算结果的解释方法，而不需要了解其具体运算过程，就可以在使用手册的帮助下完成对数据的定量分析。现在很多用户只需要适当练习，就能掌握简单的操作分析，因此 SPSS 特别受非统计专业数据分析人员的青睐。

SPSS 采用类似 Excel 表格的方式输入与管理数据，数据接口较为通用，能方便地从其他数据库中读入数据，包括常用的、较为成熟的统计方法，完全可以满足非统计专业人士的工作需要，是非统计专业人员的首选统计软件。

3. 功能全面

SPSS 对初学者、熟练者及精通者都比较适用，提供了数据获取、数据管理与准备、数据分析、结果报告这样一个数据分析的完整过程，因此非常全面地涵盖了数据分析的整个流程，特别适合设计调查方案、对数据进行统计分析，以及制作研究报告中的相关图表。

此外，SPSS 具有完整的数据输入、编辑、统计分析、报表、图形制作等功能。仅 SPSS Base 模块就提供了从简单的统计描述到复杂的多因素统计分析方法，如数据的探索性分析、统计描述、列联表分析、二维相关、秩相关、偏相关、一元方差分析、非参数检验、多元回归、生存分析、协方差分析、判别分析、因子分析、聚类分析等常见的分析方法。

4. 强大的编程功能，支持二次开发

对于常见的统计方法，SPSS 的命令语句、子命令及选择项的选择等绝大部分采用“对话框”

的操作方式。因此，用户无须花大量时间记忆大量的命令、过程、选择项。

由于 SPSS 23 具备强大的 Syntax 编程功能，且包括了 SPSS Programmability Extension 模块的编程扩展功能，那些熟练或精通者也较喜欢 SPSS，因为他们可以通过编程，在 SPSS 命令语法语言的基础上提供与其他编程语言的结合功能，来实现更强大的功能。例如，用其他语言编写的程序代码，如 Python 和 R，可以管理使用 SPSS 语法编写的任务流。使用 SPSS 23 提供的扩展编程功能和特性，让 SPSS for Windows 成为了最强大的统计开发平台之一。

5. 支持丰富的数据源，具备强大的数据访问和管理能力

SPSS 可以同时打开多个数据集，方便研究时对不同数据库进行比较分析和数据库转换处理。软件提供了更强大的数据管理功能，可帮助用户通过 SPSS 使用其他应用程序和数据库。能够读取及输出多种格式的文件，比如，由 dBase、FoxBase、FoxPro 产生的*.dbf 文件，文本编辑器软件生成的 ASCII 数据文件，Excel 的*.xls 文件等均可转换成可供分析的 SPSS 数据文件。支持 Excel、文本、dBase、Access、SAS 等格式的数据文件。能够把 SPSS 的图形转换为 7 种图形文件。结果可保存为*.txt、Word、PPT 及 Html 格式的文件。

此外，通过使用 ODBC 的数据接口，可以直接访问以结构化查询语言（SQL）为数据访问标准的数据库管理系统，通过数据库导出向导功能可以方便地将数据写入到数据库中，等等。

在 SPSS Statistics 23 中，新增加了 Salesforce.com 的数据库驱动程序，允许分析人员访问 Salesforce.com 中的数据，就像访问 SQL 数据库中的数据一样。

6. 灵活的配置方案

对于商业用户而言，SPSS Statistics 是一种按照模块进行配置的软件产品，主要包括 SPSS Statistics Base 模块和其他一系列扩充功能模块。SPSS Statistics Base 是基础的软件平台，具备强大的数据管理能力和输入输出界面管理能力，并具备完备的常见统计分析功能，而其他每个独立扩充功能模块均在 SPSS Statistics Base 的基础上，为其增加某方面的分析功能。

7. 支持多种操作系统

客户端 SPSS 支持 Windows（32 位和 64 位）、Linux 和 Mac OS。服务器端 SPSS 支持 Windows Server 2003（32 位和 64 位）、Windows Server 2008（32 位和 64 位）、AIX、HP-UX、Solaris。

1.1.3 SPSS 23 的新特性

IBM SPSS 公司最新发行的 SPSS Statistics 23，保留了 SPSS Statistics 软件一贯的界面风格，并在原先版本基础之上进一步完善了分析功能，提高了分析性能和交互能力，新版本软件有如下增强和改进。

1. 地理空间关联规则

通过使用地理空间关联规则，可以根据空间属性和非空间属性在数据中查找模式。例如，可以通过位置属性和人口统计信息属性识别罪案数据中的模式。根据这些模式，可以构建规则，以便预测有可能发生特定类型罪案的地点。

2. 空间时间预测

空间时间预测使用包含位置数据、预测输入字段（预测变量）、时间字段和目标字段的数据。每个位置在数据中都有许多行，这些行表示每个预测变量在每个位置与每个时间间隔的值。

3. 时间因果模型

时间因果建模尝试发现时间序列数据中的关键因果关系。在时间因果建模中，可指定一组目

标序列以及这些目标的候选输入集，这样，过程将为每个目标构建一个自回归时间序列模型，并且仅包括那些与目标具有因果关系的输入。此方法不同于传统时间序列建模，在传统时间序列建模中，必须为目标序列显式指定预测变量。由于时间因果建模通常涉及为多个相关的时间序列构建模型，因此结果称为模型系统。

4. 批量装入数据库

将数据导出至数据库时，批量装入会将数据成批提交到数据库，而不是一次提交一条记录。此操作可以使操作速度更快，对于大型数据文件尤其如此。

5. 可编程性增强功能

现在，可以从任何外部 R 进程运行使用 R Integration Package for IBM® SPSS® Statistics 中的函数的 R 程序，如 R IDE 或 R 解释器。还可以从 R 运行 SPSS Statistics 命令语法。

通过 Python 或 R 实现的扩展命令现在支持在变量列表中使用 TO 和 ALL 关键字。IBM SPSS Statistics-Essentials for R 和 IBM SPSS Statistics-Essentials for Python 现在包含更多扩展命令以及关联的定制对话框。另外，可以通过在语法编辑器中按 F1 键来访问随 Essentials for R 和 Essentials for Python 一起安装的所有扩展命令的帮助。

6. 更快的执行性能

建立数据透视表的输出速度比以前提升了两倍。因此，当涉及大型报表输出或大量运算的报表时，可大大节省做报告的时间。此外，输出报表所占用的内存空间也将更少。

1.1.4 SPSS 的模块

SPSS 统计分析软件是一款模块化设计的产品，用户可以根据需要选择功能模块进行配置购买，以节省费用。它主要包括 SPSS Statistics Base 模块和其他一系列扩充功能模块，共 16 个，每个独立扩充功能模块均可在 SPSS Statistics Base 模块基础上，为其增加某方面的分析功能。下面简要介绍 16 个模块的功能。

1. Statistics Base 模块

Statistics Base 模块管理整个软件平台，以及数据访问、数据处理和输出，并能进行很多种常见的基本统计分析和报告，其中包括计数、交叉表和描述统计、OLAP 立方和码本报告。它还提供了多种降维、分类和细分方法，例如因子分析、聚类分析、最近邻元素分析和判别函数分析。此外，SPSS Statistics Base 模块还提供了广泛的平均值比较算法和预测方法，例如 T 检验、方差分析、线性回归和序数回归。

2. Advanced Statistics 模块

Advanced Statistics 模块为顺序结果建立更灵活、更成熟的模型，在处理嵌套数据时可得到更精确的预测模型，可以分析事件历史和持续时间数据。它包括广义线性模型（GZLMS）、广义估计方程（GEES）、混合模型、一般线性模型（GLM）、方差成分估计、MANOVA、Kaplan-Meier 估计、Cox 回归、多因子系统模式的对数线性模型、对数线性模型、生存分析。

3. Bootstrapping 模块

Bootstrapping 模块可以更有效地使用小样本量的数据，通过数据自身重采用的功能，可以模拟大样本情况下的采样结果，从而对数据结构特征和偏差有更直接的认识。该方法可以导出稳健

的标准误估计值，并能为诸如均值、中位数、比例、比率、相关系数或回归系数等估计值导出置信区间。

4. Categories 模块

Categories 模块用启发性的二维图和感知图清晰地表现数据中的关系，可以更完整、更方便地分析数据。通过启发性的概念映射、最优尺度、偏好尺度和数据降维技术，揭示数据中全部潜在的关系。

5. Complex Samples 模块

如果使用了特别复杂的抽样方案，该模块可以计算复杂样本的统计数据，得到更精确的结果。它拥有专门的规划工具和统计方法，提供各种向导来制定取样方案或详细定义样本，并提供专门的技术来解决样本设计及相伴标准误差，能够减少得出错误或误导性推论的风险。**Complex Samples** 模块将抽样设计融入调查分析之中，对复杂抽样数据的总体得到更加有效的统计推论，对于调查、市场、民意研究人员或者社会科学家来说是必不可少的统计工具。

6. Conjoint 模块

Conjoint 模块提供一种实际的方式，用以度量单个产品属性如何影响消费者和市民偏好。帮助市场研究人员和新产品开发部门了解在消费者心目中什么产品属性是重要的，了解最偏爱的属性水平是什么，从而进行定价研究，以及品牌价格研究。在产品投入大批量生产之前进行这些研究，以避免可能的失误。

7. Custom Tables 模块

Custom Tables 模块可帮助分析人员在较少的时间里创建各种美观、精确的表格，包括复杂的行列表和多重响应数据的显示。

8. Data Preparation 模块

Data Preparation 模块可以简化数据准备过程，在预处理数据时轻易地识别虚假的和无效的观测、变量和数据值，确认可疑的或者残缺的案例，查看数据缺失模式，描述变量分布以备分析，更准确地应用针对于分类变量的算法，还可以用为分类变量设计的运算法则来做更多精确的工作。使用 **Data Preparation** 模块，可以迅速找到多元的极端值，执行数据检验，为建模预处理数据。

9. Decision Trees 模块

Decision Trees 模块基于数据挖掘中发展起来的树结构模型对分类变量或连续变量进行预测，可以方便、快速地对样本进行细分。此过程为探索性和证实性分类分析提供验证工具。

10. Direct Marketing 模块

Direct Marketing 模块主要用来处理市场直销中的一些分析需求。目前提供 **RFM** 客户评分、客户分群、目标客户轮廓概括、客户响应评分、不同营销行为响应测量等模型，使其营销计划尽可能地发挥效力。

11. Exact Tests 模块

Exact Tests 模块可在小样本或分布非常不均匀的样本可能导致常规检验不准确的情况下计算统计检验的精确 P 值。此选项只在 **Windows** 操作系统中可用。

12. Forecasting 模块

Forecasting 模块通过使用多种曲线拟合模型、平滑模型和用于估计自回归函数的方法，执行综合的预测和时间序列分析。

13. Missing Values 模块

缺失数据会带来偏差或错误的分析结果，简单代入法或者简单的回归法都不能正确地填补缺失值，Missing Values 模块描述了缺失数据的模式、估计均值和其他统计量，并利用统计方法填充缺失值。

14. Neural Networks 模块

Neural Networks 模块可以通过将产品需求预测为价格函数以及其他变量的函数，或根据购买习惯和人口统计特征对客户进行分类来制定经营决策，是非线性数据建模工具。它们可以用来建立输入与输出之间的复杂关系模型，也可用来查找数据中的模式。

15. Regression 模块

Regression 模块提供了用于分析不能拟合传统线性统计模型的数据的方法。它包括一些用于 probit 分析、logistic 回归、权重估计、两阶段最小平方回归和常规非线性回归的过程。

16. Programmability Extension 模块

Programmability Extension 模块可以使用外部语言来执行 SPSS 一连串的分析动作，以达到自动化的目的。通过撰写内嵌在 SPSS 23 里的 Python 程序来控制 SPSS 的各式语法工作执行，如执行设定变量属性、观察程序输出、错误码或条件状态等。运用外部程序结合新的后端 API 处理，可扩大语法执行工作的弹性。

1.2 SPSS 使用基础

1.2.1 SPSS 的安装

IBM SPSS Statistics 23 for Windows 的安装同其他 Windows 应用软件一样，非常简单，下面简要介绍安装 SPSS 的步骤。

1. 启动安装

将 SPSS 软件安装盘放入光驱，如果系统设置为自动运行光盘状态，则光盘自动执行 setup 应用程序，出现安装界面，如图 1-1 所示。若不能自动运行，则运行资源管理器，打开光盘中的 Windows\setup.exe 文件，出现安装界面（注：如果是 SPSS 官方网站下载的试用版，则直接运行安装文件，进入安装过程）。

2. 设置安装选项

按照安装向导，根据提示设置安装信息：

- 正版 SPSS 需输入 SN 序列号，试用版不需要。
- 安装过程中会询问许可证的不同类型，即用户的不同类型，如图 1-2 所示。根据实际购买情况，如果是单机用户，选择选项“单个用户许可证”；如果企业购买的软件是网络版，则选择选项“网络许可证”。

- 需要接受软件使用协议。
- 用户需要填写用户名、单位名称，如果是网络版，需设置许可证服务器名称或地址。
- 选择帮助语言，默认是“英文”+“中文”，试用版需单独下载帮助语言包。
- 系统默认的安装路径是 C:\Program Files\IBM\SPSS\Statistics\23\，如用户需改变安装路径，可以单击“更改”按钮来自定义安装位置。
- 单击“安装”按钮，开始安装软件。

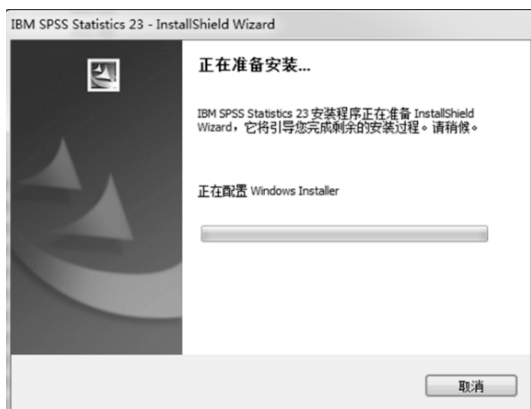


图 1-1 安装启动界面



图 1-2 许可证类型选择界面

3. 软件授权

安装完毕，启动 SPSS 授权过程，根据软件的授权码，连接 SPSS 公司的许可证管理服务器获取许可证。成功授权之后，软件方可正常使用。试用版有临时许可证供短期试用。

1.2.2 SPSS 的界面

SPSS 主要有 5 类窗口，分别为数据编辑窗口、结果输出窗口、结果编辑窗口、语法编辑窗口和脚本编辑窗口。

1. 数据编辑窗口

数据编辑窗口是 SPSS 软件中最常用的窗口，这个窗口主要用来处理数据和定义数据字典，它分为两个视图：一个是用于显示和处理数据的数据视图（Data View），另一个是用于变量定义和查看的变量视图（Variable View）。

数据视图如图 1-3 所示，提供类似 Excel 电子表格的编辑窗口，在该窗口中可以创建、编辑、浏览数据文件。其操作和 Excel 非常相似。在 SPSS 中允许打开多个数据文件名进行编辑、浏览，正在编辑的数据文件称为活动数据文件，只有活动数据文件的数据才能被分析处理。SPSS 的数据以表格的形式呈现，表的每一行表示一个观察个案（Case），每一列表示一个变量（Variable），表的大小由变量数和观察个案数确定。一般情况下，分析的数据应以 SPSS 数据文件的形式保存，最常用的 SPSS 数据文件扩展名为“*.sav”，保存数据文件的同时也保存了变量属性和变量值。

变量视图的功能是定义数据集的数据字典，它用来定义、显示和修改数据集中的变量信息，变量视图如图 1-4 所示。

SPSS 的功能主要通过菜单和工具栏实现，工具栏是常用菜单项的快捷方式，下面介绍菜单的主要功能。



图 1-3 数据视图



图 1-4 变量视图

- (1) 文件：“文件”菜单负责新建各种类型的文件、读入文件或数据库的内容、保存文件、将数据输出到数据库、标记文件为只读文件、重命名数据集、打印等操作。其中需要特别指出两个功能：一个是缓存数据，它可以将数据载入内存，大大提高运行速度；另一个是开关服务器，可以连接安装有 SPSS 服务器版本的高性能服务器，进行分布式分析。
- (2) 编辑：“编辑”菜单对文件数据进行选择、复制、粘贴、删除、查找，还可以插入变量、个案，选择“选项”可以进行 SPSS 的常规、编辑、格式等全局选项设置。
- (3) 查看：“查看”菜单对软件界面的显示进行修改，可以显示或隐藏状态栏，添加工具栏，编辑菜单栏，进行字体设置，显示或隐藏值标签。
- (4) 数据：“数据”菜单，进行数据变量的定义、复制数据或数据集、定位观测量、分类观测量、转换重构变量、合并拆分文件、数据异常检查及加权等操作。
- (5) 转换：“转换”菜单，进行数值的计算、重新编码、离散化处理、缺失值替代、创建时间序列、产生随机数等操作。
- (6) 分析：“分析”菜单，应用各种统计方法对当前窗口中的数据进行分析，包含了 SPSS 的核心功能。
- (7) 直销：“直销”菜单主要用来处理市场直销中的一些分析需求。
- (8) 图形：“图形”菜单根据当前数据绘制和编辑各种统计图表，如条形图、散点图、线图、面积图、直方图、箱图、饼图等。
- (9) 实用程序：“实用程序”菜单进行变量列表、控制输出管理系统、输出文件信息、定义和使用变量集合、生产设施、集成 R 或者 Python 的外部程序等操作。
- (10) 窗口：“窗口”菜单，进行窗口拆分、最小化、切换窗口等操作。
- (11) 帮助：“帮助”菜单，提供 SPSS 系统帮助、教程、个案研究、统计辅导、指令语法及算法参考等。

2. 结果输出窗口

SPSS 的结果输出窗口也称为结果视图或者结果浏览窗口，该窗口用于存放 SPSS 的操作日志及分析结果，如图 1-5 所示。整个窗口分为两个区：左边为目录区，是 SPSS 分析结果的目录；右边是内容区，显示与目录对应的内容。在结果浏览窗口内可以浏览、编辑输出结果，改变输出显示顺序等。

SPSS 的结果输出可以保存为“*.SPV”的文件格式，还可以将全部或选定部分结果导出为 Html、Word、PPT、PDF 等多种格式的文件。

3. 结果编辑窗口

结果编辑窗口是编辑分析结果的窗口。在结果视图中，选择要编辑的内容，双击或单击右键选择“编辑内容”，选中的图表可以在单独的窗口中进行编辑，对于表格还可以直接在结果窗口中编辑。如图 1-6 所示为图形的结果编辑窗口。

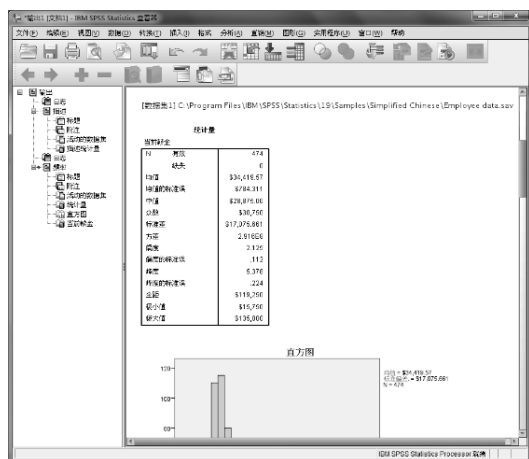


图 1-5 结果输出窗口

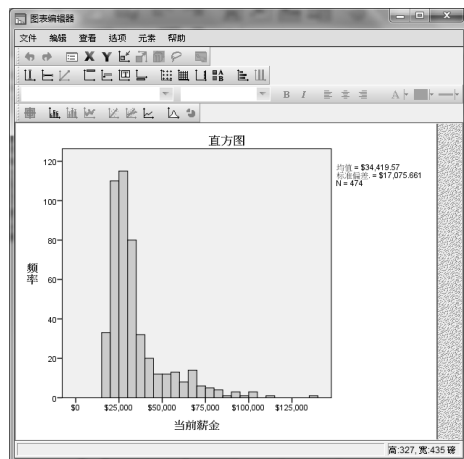


图 1-6 结果编辑窗口

4. 语法编辑窗口

SPSS 除了提供菜单操作外，还提供语法编程方式，可通过“文件→新建→语法”新建一个 SPSS 语法程序，也可通过“文件→打开→语法”打开一个 SPSS 语法文件。语法编程除了能够完成窗口操作所能完成的所有任务外，还能完成许多窗口操作所不能完成的其他工作，实现分析和控制自动化。语法编辑窗口是编写、调试和运行 SPSS 程序的窗口，如图 1-7 所示。

5. 脚本编辑窗口

在 SPSS 数据编辑窗口或结果浏览窗口中执行“文件→新建→脚本”命令，出现如图 1-8 所示的宏程序编辑窗口，在该窗口中可以用 VB 语言编程，实现用户特殊的需要。

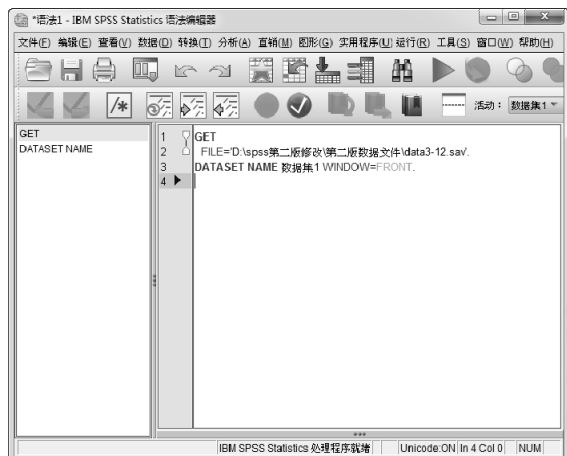


图 1-7 语法编辑窗口

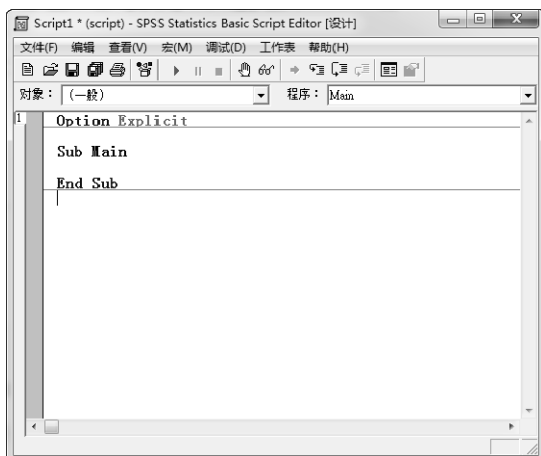


图 1-8 脚本编辑窗口

1.3 SPSS 的帮助系统

SPSS 提供了功能全面的在线帮助系统，是我们学习和使用 SPSS 最权威的参考。SPSS 提供的帮助系统包括主题、教程、个案研究、统计辅导、指令语法参考、算法等，另外，SPSS 系统的每个对话框都提供联机帮助。SPSS 23 与之前的版本不同，它的帮助系统采用 Html 网页形式，通过浏览器进行查阅。

1.3.1 主题

按主题组织的帮助系统提供全部 SPSS 菜单操作和相关内容的帮助，有索引，可按关键词搜索，是 SPSS 的主要在线帮助功能。

选择菜单“帮助→主题”，启动默认浏览器，出现如图 1-9 所示的主题帮助窗口。该窗口分为两个子窗口，左边导航窗口显示查找信息的主题目录，右侧内容窗口显示具体帮助内容。左侧导航窗口可以按 4 种方式搜索浏览相应的信息，即目录浏览方式、关键词索引浏览方式、关键词搜索方式和用户自定义的书签。可以单击左侧窗口下面相应的选项卡来切换不同的浏览方式。

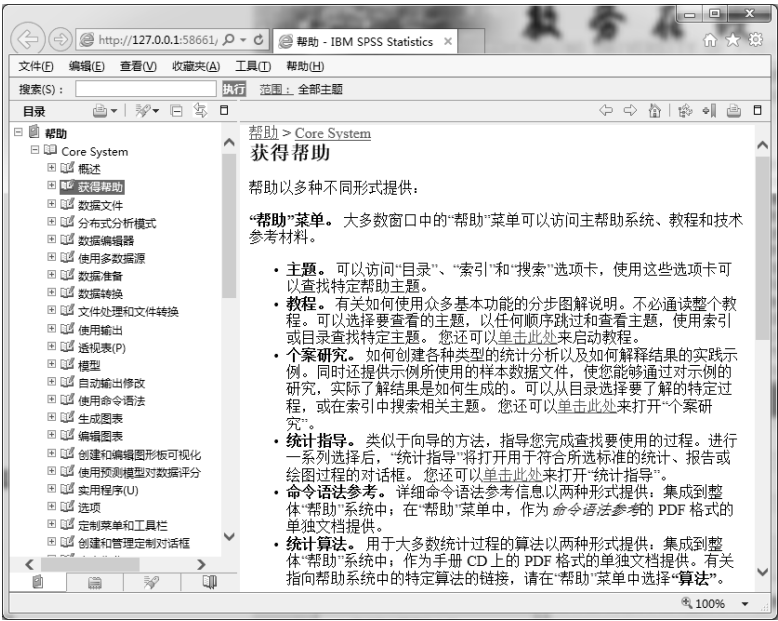


图 1-9 主题帮助窗口

1.3.2 教程

SPSS 教程是为初学者提供的学习资料，以动画的形式一步步演示和介绍 SPSS 中的基本操作步骤，如图 1-10 所示。在该窗口中，中部显示的是具体操作视图，右侧是具体操作步骤及解释，右下方的两个箭头表示向前翻页和向后翻页，右上角的 7 个按钮从左到右分别表示后退、前进、主页、在目录中显示、添加书签、打印及最大/最小化。

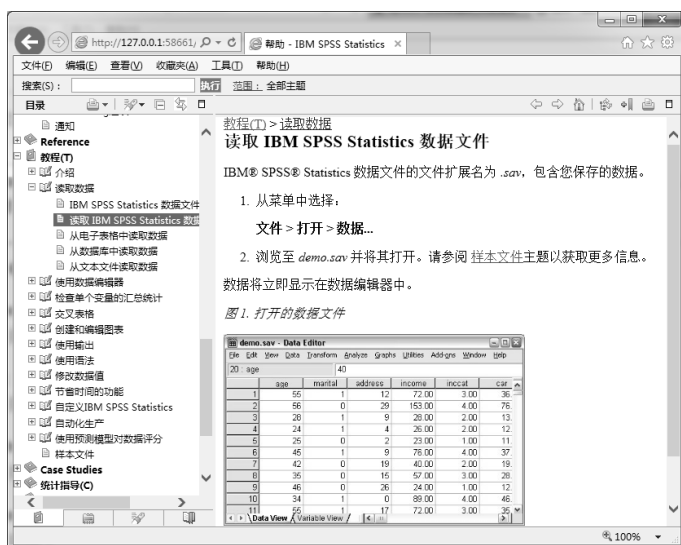


图 1-10 教程

1.3.3 个案研究

最快的学习方法是通过对实例进行具体操作，SPSS 中的个案研究（Case Studies）帮助模块就是通过向用户展示具体的实例操作过程来帮助用户尽快掌握 SPSS 的功能，如图 1-11 所示。SPSS 23 中的个案研究没有中文版的，该窗口同 1.3.2 节的教程帮助窗口类似，都是以动画演示配合操作步骤描述的形式给出帮助。

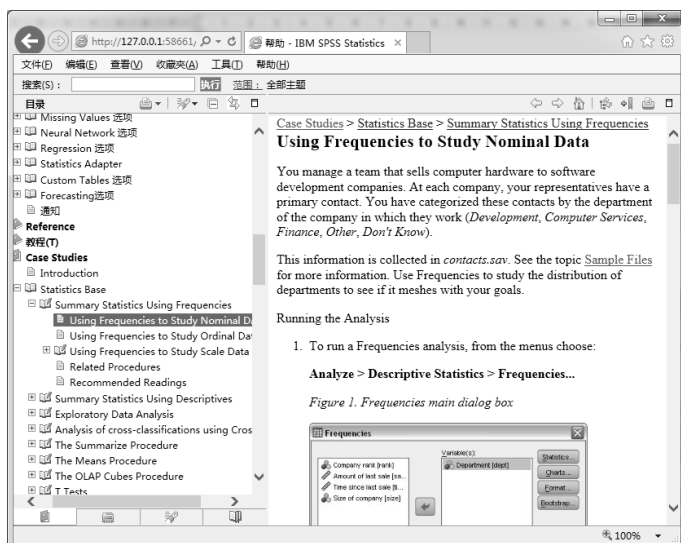


图 1-11 个案研究

1.3.4 统计辅导

统计辅导的目的是指导用户找到并使用正确的 SPSS 过程来进行统计分析，如图 1-12 所示。在窗体中询问用户想要通过 SPSS 实现什么功能，选择后会给出实现该功能的方法和操作步骤。

1.3.5 高级帮助

对于统计专业人士和 SPSS 的高级用户，SPSS 软件在帮助菜单中给出了两个不同内容的帮助文档，即指令语法参考和算法。另外，SPSS 帮助中还有一个开发者中心，指向 SPSS 社区网站，供 SPSS 各级用户及开发者交流及共享资源。

1. 指令语法参考

选择菜单“帮助→指令语法参考”，打开如图 1-13 所示的指令语法参考 PDF 帮助文档。该帮助文档提供了 SPSS 中所有命令语法以及相应的示例，并对 SPSS 命令语法给出了详细的解释，供利用 SPSS 命令语法程序进行数据管理和统计分析的专业人士参考。

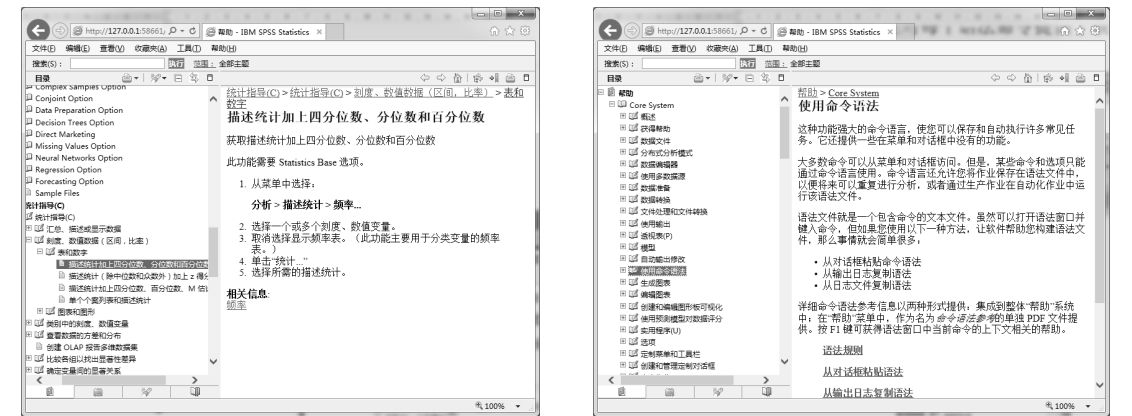


图 1-12 统计辅导

图 1-13 指令语法参考

同时，SPSS 也对每个语法命令提供了联机帮助，在语法命令编辑器的相应语法命令中，按 F1 键即可提供该语法命令的联机帮助。

2. 算法

选择菜单“帮助→算法”，出现如图 1-14 所示的 SPSS 算法帮助文档。算法帮助文档提供了 SPSS 各种统计分析所采用的算法，供进行理论分析的专业人士参考。

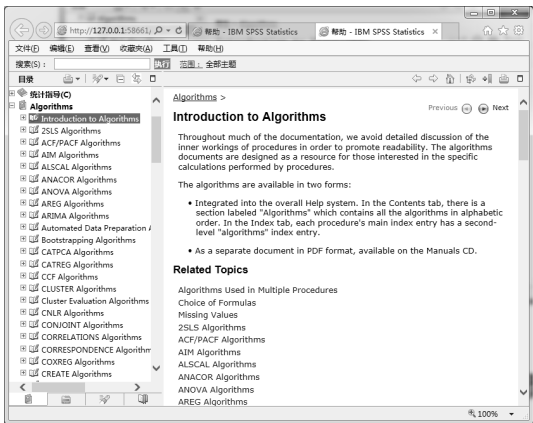


图 1-14 算法帮助文档

1.4 利用 SPSS 进行数据分析的步骤

学习和应用 SPSS 软件的过程并不是单纯地学习和应用一种计算机软件的过程。由于 SPSS 是一种专业性较强的统计软件，因此，学习和应用它时必须了解和掌握必要的统计学专业知识，以及数据分析的一般步骤和原则，这样才能避免滥用和误用方法，避免因偏差甚至错误的数据分析结论而做出错误的决策。

1.4.1 统计学中数据分析的一般步骤

统计学通常被定义为用以收集数据、分析数据和由数据得出结论的一组概念、原则和方法，因此，在数据分析的实践中，用统计学的理论指导应用是必不可少的，也是极为重要的。数据分析一般经过收集数据、加工和整理数据、分析数据等几个主要阶段。

1. 明确数据分析目标

明确数据分析目标是数据分析的出发点。明确数据分析目标就是要明确本次数据分析要研究的主要问题和预期的分析目标等，只有明确了数据分析的目标，才能正确地制定数据采集方案，例如，应收集哪些数据？应采用怎样的方法收集？等等，进而为数据分析做好准备。

2. 正确收集数据

正确收集数据是指应从分析目标出发，排除干扰因素，正确收集服务于既定分析目标的数据。正确的数据对于实现数据分析目标起到关键性的作用。在收集数据的过程中，经常会获得一些与分析目标无关或者对分析目标起相反作用的干扰数据，排除这些数据是数据收集的重要环节。数据分析并不仅仅是对数据进行数学建模，收集的数据是否真正迎合数据分析的目标，其中是否包含其他因素的影响，其影响程度如何，应如何剔除这些影响等问题都是数据分析中必须注意的重要问题。

3. 数据的加工整理

在明确数据分析目标的基础上收集到的数据，往往还需进行必要的加工整理后才能真正用于分析建模。通过数据的加工整理，人们能够大致掌握数据的总体分布特征，这是今后进一步深入分析和建模的基础。数据的加工整理通常包括数据的缺失值处理、数据的分组、基本描述统计量的计算、基本统计图形的绘制、数据取值的转换、数据的正态化处理等。

4. 明确统计方法的含义和适用范围

数据加工整理完成后，一般就可做进一步的数据分析了。分析时应切忌滥用和误用统计分析方法。滥用和误用统计分析方法主要是由于对方法能解决哪类问题、方法适用的前提、方法对数据的要求不清等原因造成的。另外，统计软件的不断普及以及应用中的不求甚解也会加重这种现象。在明确了统计方法的含义和适用范围后，选择几种统计分析方法对数据进行探索性的反复分析也是极为重要的。每一种统计分析方法都有自己的特点和局限性，因此一般需要选择几种方法反复进行分析，仅依据一种分析方法的结果就断然做出结论是不科学的。

5. 正确解释分析结果

数据分析的直接结果是统计指标和统计参数。正确理解这些指标和参数的统计含义是一切分析结论的基础，不仅能够帮助人们有效避免毫无根据地随意引用统计数据的错误，同时也是证实分

析结论正确性和可信性的依据。而这一切都取决于人们能否正确地把握统计分析方法的核心思想。

另外，将统计指标和统计参数与实际问题相结合也是非常重要的。客观地说，统计方法仅仅是一种有用的数量分析工具，它绝不是万能的。统计方法是否能够正确地解决各学科的具体问题，不仅取决于应用统计方法、工具或人能否正确地选择统计方法，还取决于它们是否具有深厚的应用背景。只有将各学科的专业知识与统计指标和统计参数相结合，才能得出令人满意的分析结论。

1.4.2 利用 SPSS 进行数据分析的一般步骤

利用 SPSS 进行数据分析也应遵循数据分析的一般步骤，但涉及的方面会相对较少，主要的工作集中在以下 4 个阶段。

1. SPSS 数据的准备阶段

按照 SPSS 的要求，利用 SPSS 提供的功能准备 SPSS 数据文件，其中包括在数据编辑窗口中定义数据的结构、录入和修改数据等。

2. SPSS 数据的加工整理阶段

对数据编辑窗口中的数据进行必要的预处理。例如，将数据的缺失值补齐，对数据进行排序、拆分等。

3. SPSS 数据的分析阶段

选择正确的统计分析方法，对数据编辑窗口中的数据进行分析建模。由于 SPSS 能够自动完成数据建模中的数学计算并给出计算结果，使分析人员无须记忆数学公式，这无疑给统计分析和 SPSS 的广泛应用铺平了道路。

4. SPSS 分析结果的解释

读懂 SPSS 输出窗口中的分析结果，明确其统计含义，并结合应用背景知识做出切合实际的合理解释。在以后的章节中会介绍分析实例中结果的统计含义，便于用户更好地理解分析结果。

第 2 章 统计数据的收集与预处理

如果想了解某一具体经济领域或者企业内外部的情况，以及某一具体目标市场的情况，获取准确的第二手数据是相当困难的，此时我们一般采用的方法是针对具体的目标进行市场调查，获得第一手数据，并基于此数据进行分析，得出结论。市场调查是获得第一手数据的主要手段，其涉及的知识相当多，从调查问卷的设计、调查方法的选择、抽样方案的确定、抽样实施、问卷回收数据录入到最后的数据分析并得出结论是一个整体，本书限于篇幅，不能对所有的流程进行详细说明，只针对其中最重要的两个环节（问卷设计、问卷分析）做简要阐述。

通过市场调查获得第一手数据后，需要对数据进行编码、录入和整理，这是研究者利用 SPSS 进行统计分析的必要前提，只有准确地建立了高质量的数据文件，才能保证数据分析结果的正确性和科学性。本章主要介绍在进行统计分析之前，如何通过市场调查获得统计数据，获得数据后，如何将问卷调查获得的资料转变为 SPSS 能够识别、统计的数据文件，并对数据文件做必要的预处理，为进一步的统计分析做好准备。

2.1 统计数据的收集

2.1.1 问卷设计

对于问卷的设计，主要是在清楚调查内容的基础上，通过对调查问题的合理设计和布局，让问题更好地反映调查内容。一项以一手数据为基础的研究项目，其研究深度本质上由问卷的深度决定。问卷设计时没有考虑到的问题，在问卷调查完成后再想研究就不太可能了，因为重新设计问卷、收集数据，无论是时间还是资金的损失，往往都是难以承受的。因此，按照所要研究的问题设计好问卷，是科学研究中一项非常重要的工作，下面就问卷的构成、问卷的问题类型、问卷中量表的主要类型和问卷设计的注意事项几个方面进行阐述。

1. 问卷的构成

一份正式的调查问卷一般包括以下 4 个组成部分：标题、导语（前言）、正文和结束语。

（1）标题

问卷的标题概括地说明调研主题，使被访者对所要回答的问题有一个大致的了解。问卷标题要简明扼要，但又必须点明调研对象或调研主题，如“学生宿舍卫生间热水供应现状的调研”，而不要简单采用“热水问题调查问卷”这样的标题，否则无法使被访者了解明确的主题内容，妨碍回答问题的思路。

（2）导语（前言）

导语（前言）主要是对调查目的、意义及填表要求等的说明，导语部分文字须简明易懂，能激发被调查者的兴趣。导语的书写应注意以下 3 个问题：

- 与调查的内容和调查对象相吻合，突出本次调查的主要问题和现象。
- 要调动被调查者的积极性，体现被调查者完成本次调查的重要作用。
- 最后要对被调查者表示感谢，写一些祝愿的话语。

另外，卷首最好要有说明（称呼、目的、填写者受益情况、主办单位），如涉及个人资料，应该有隐私保护说明。

(3) 正文

将调查的若干问题及相应的选择项目有限度地排列，要求被调查者回答。

(4) 结束语

一般是一段短语，内容是向被调查者的合作再次表示感谢，以及关于不要漏填及复核的请求。结束语要简短明了，有的问卷也可以省略。

2. 问卷的问题类型

(1) 封闭型问题

封闭型问题事先准备了答案，应答者只能在事先准备的答案中选择。封闭型问题的数据转化工作量大为减少。例如“我在英语课堂活动的参与程度：A. 十分活跃 B. 较活跃 C. 一般 D. 不太活跃 E. 根本不参与”，这是固定应答题，对指定答案方式进行回答。

封闭型问题包括以下问题形式：

➤ 是否式

问题后列出两种相互对立的答案，从中选出一个，“是”与“否”，或“同意”与“不同意”。例如“我觉得英语是一门十分有趣且很好学的学科。A. 同意 B. 不同意”。

➤ 选择式

根据自己的观点和实际情况，从列出的多个答案中挑选出一个或几个答案。例如“我在课后对课堂所学知识的整理：A. 十分认真 B. 较认真 C. 一般 D. 不太认真 E. 一点也不认真”；“我觉得英语学习中最重要的是（最多选三项）：A. 单词记忆 B. 语法规则 C. 阅读理解 D. 口语表述 E. 写作能力 F. 听力 G. 做题技巧 H. 语言知识”。

➤ 评判式

评判式也叫排列式、编序式，每个问题后列出很多选项，根据重要性用数字评定等次。例如“英语学习涉及很多内容，请根据你的观点，按其重要程度由 1 到 5 顺序排列。

() 听力技能 () 口语表达技能 () 阅读技能 () 写作技能 () 翻译技能”。

在多数情况下，要尽量用封闭型问题形成问卷。

(2) 开放型问题

开放型问题事先没有准备答案，允许被调查者用自己的话来回答问题，由于采取这种方式提问会得到各种不同的答案，不利于资料的统计，通常在问卷形成阶段使用，在最终问卷中要慎用。

3. 问卷中量表的主要类型

量表是测量应答者对某个问题（特别是复合型问题）的反应强度（或态度、看法）的工具。把单选问题的备选答案量化，就得到单问题量表。如表 2.1 所示。

表 2.1 单问题量表

金融危机对你的影响	非常大	大	一般	不大	无影响
	5	4	3	2	1

这就是一个简单的单项量表。“单项”是指该量表仅仅反映了应答者对一个问题的态度。

(1) 连续评分量表

上述量表的刻度仅从 1 到 5，如果采用 0 到 100 的刻度，则称为连续评分量表。

(2) 分项评分量表（Likert 量表）

上述量表只涉及一个单选问题，如果量表涉及多个关联的单选问题，就称为分项评分量表

(Itemized Rating Scale)。分项评分量表中的多个单选问题必须有关联，是某个总项（上一层的变量）的一个分解。表 2.2 就是一个分项量表的例子。

表 2.2 高校辅导员压力问题的一个分项评分量表

项 目	没有压力	有点压力	中度压力	压力较大	压力很大
辅导员外出进修和接受继续教育的机会	1	2	3	4	5
辅导员工作权利和义务不明确	1	2	3	4	5
事务性工作多，用于自身学习提高的时间少	1	2	3	4	5
辅导员工作见效周期长，成果无形化	1	2	3	4	5
辅导员评职称的条件及难度	1	2	3	4	5

这种分项评分量表又称 Likert 量表，是由美国社会心理学家 Rensis A.Likert 于 1932 年提出的。Likert 量表的度量级别，通常是 5 级，但不一定是 5 级，在应用中 7 级、9 级均可，但通常不少于 5 级，不高于 9 级。

Likert 量表的关键特点是所有的“分项”共同组成一个总项，分项的得分加总后就得到总项的得分，所以 Likert 量表又称为加总量表（或求和量表）。

(3) 排序量表

排序量表就是根据所研究的问题对几个因素排列出先后顺序，例如，根据重要性对影响学生成绩的因素进行排序的量表如表 2.3 所示。

表 2.3 对学生成绩影响的重要性等级表

比较因素	重要性等级
学生智商	5
家长监管	4
学习勤奋程度	3
教学质量	2
社会因素	1

4. 问卷设计的注意事项

(1) 目的明确

任何问卷调查都是有目的的，要么证实某个结论，要么证伪某个结论，只有目的明确，才能围绕目的设计题项。

(2) 先易后难，先简后繁

问卷头几个问题的设置必须谨慎，招呼语措辞要亲切、真诚，前面几个问题要比较容易回答，不要使对方难以启齿，给接下来的调研造成困难，所以通常将被调查者熟悉的问题放在问卷的前面。

(3) 提出的问题要具体，避免提一般性问题

一般性的问题对实际调研工作并无指导意义，例如，“你认为食堂的饭菜供应怎么样？”这样的问题就很不具体，很难达到想了解被访者对食堂饭菜供应状况总体印象的预期调研效果，应把这一类问题细化为具体询问关于价格、外观、卫生、服务质量等方面的印象。

(4) 单选问题的备选答案应完整划分答案空间。

单选问题的备选答案必须分布在同一个维度上，是同一个答案空间的完整分割，即备选答案之间不能有交集，也不能有遗漏。例如，如下问题的 5 个备选答案就是一个答案空间的完整划分。

“您的家庭月收入是：

- A. 2000 元以下
- B. 2000~3999 元
- C. 4000~5999 元
- D. 6000~7999 元
- E. 8000 元及 8000 元以上”

收入的所有答案都可以在这 5 个备选答案中找到自己的位置，是一个答案空间的完整划分，没有交集，也没有遗漏。

(5) 多选题的备选答案必须分布在两个以上的维度上，并且至少有一部分不是互相排斥的。

- (6) 问题的陈述及备选答案不能有多重含义。
- (7) 问题设计的用语要含义明确,不能让应答者产生不同的理解。
- (8) 在问题的陈述中,要对所询问行为的时间、方式、目的做必要的限定。
- (9) 对于得不到诚实回答而又必须了解的数据,可以通过变换问题的提法来获得相应的数据,或者通过了解相对数据来判断总体的情况。
- (10) 问卷不能太长,以 20~30 分钟为宜;商场拦截类的问卷,以 3~5 分钟为宜。

2.1.2 问卷分析

在问卷调查过程中,通过问卷得到的调查结果与真实情况之间不可避免地会存在着误差,这些误差有可能是调查过程中的测量误差,也有相当一部分是由于问卷的结构质量不高造成的系统误差。因此,为了提高调查问卷的结构质量,在完成问卷设计以后,问卷还不能马上用于市场调查,还需要对问卷,特别是问卷中各量表的信度和效度进行评价。

问卷的信度(Reliability)是指如果重复进行测量,一个问卷产生一致性结果的程度。系统误差对信度没有不利影响,因为它们以不变的方式影响调查值,重复测量中真实值不会改变,因而系统误差也不会改变;相反,随机误差会影响信度,信度可以理解为调查值排除随机误差的程度,如果随机误差为 0,则调查是完全可信的。通过确定问卷系统变差的比例来评价信度,是通过确定用同一问卷得到的调查值之间的相关度来实现的,如果相关度高,则问卷可信,反之则问卷信度较低。对于调查问卷的信度,可以运用 SPSS 软件中的可靠性分析功能模块求解相关的系数,具体操作过程参见第 12 章信度分析的内容。

问卷的效度(Validity)可以定义为对象之间调查值的差异所能反映的对象之间真实值的差异程度,完美的效度要求没有测量误差,即系统误差和随机误差都为 0,在 SPSS 中有专门的信度分析模块,但没有效度分析模块,如果要进行效度分析,需要利用相关分析来实现,具体分析过程参见第 12 章。

2.2 数据文件的建立

2.2.1 统计数据的度量尺度

在统计学中,观测数据是在自然的未被控制的条件下观测到的数据,如社会商品零售额、消费价格指数、汽车销售额、降雨量等。而通过抽样调查,从全体研究对象中选取一部分个体组成样本,对样本的观测所得到的数据是试验数据。

无论是观测数据还是试验数据,都需要度量。统计数据是对客观现象计量的结果,按照对事物计量的精确程度,可将所采用的度量尺度由低级向高级分为名义尺度、定序尺度、间隔尺度。

1. 名义尺度

名义尺度即定类尺度,它仅是一种标志,用于区分变量的不同值,类别数据之间没有次序关系。它按照事物的某种属性对其进行平行的分类和分组。例如人口的性别、商品的名称、身份证、商店类型等。

名义尺度的特点是其值测度了事物之间的类别差,而对各类之间的其他差别却无法从中得知,所有类的地位相等,可以随意排序。另外,其计量结果可以且只能计算每一类别中各元素出现的频率。

使用名义尺度定义变量时,必须符合穷尽和互斥的原则,该级别的数据,对应的变量类型可以是数值型,也可以是字符型。

2. 定序尺度

定序尺度是对事物之间等级或顺序差别的一种测度。如考试成绩（优、良、中、差），人的身高等级（高、中、矮），学历等级（博士、硕士、学士）等。

定序尺度的特点是可以测度类别差，还可以测度次序差，但是定序尺度无法测出数据之间的准确差值，所以其计量结果只能排序，不能进行算术四则运算。例如，对学历等级，有“博士”>“硕士”>“学士”的次序，却不能进行四则运算。

该级别的数据，对应的变量类型可以是数值型，也可以是字符型。

3. 间隔尺度

间隔尺度是指变量的取值是连续的区间。这种尺度又可以分为定距尺度和定比尺度。

定距尺度是对事物类别或次序之间间距的测度。如 100 分制考试的成绩、重量、温度等。这种尺度的特点是，不仅能将事物区分为不同类型并进行排序，而且可准确指出类别之间的差距是多少。其次，定距尺度通常以自然或物理单位为计量尺度，因此测量结果往往表现为数值，可以进行加减运算。

定比尺度是指能够测度值之间比值的一种计量尺度。例如员工的月收入、企业的产值等。这种尺度的特点是其区间属于同一阶层，计量结果也表现为数值。

此外，它除了具有其他三种测量尺度的所有优点之外，还具有可计算两个测量值之间比值的特点，其与定距尺度的区别在与，它有一个固定的绝对“零点”，而定距尺度没有，因此它可以进行加、减、乘、除及其延伸运算，而定距尺度只能进行加减运算。间隔尺度级别的数据，对应的变量类型只能是数值型。

将统计数据组织到 SPSS 中进行分析时，也需要设置各数据的度量尺度，有些统计分析对数据的度量尺度有相应的要求，在后面的章节中将有相应的介绍。

2.2.2 SPSS 数据文件的特点

统计数据被收集以后，它们通常以典型的表格形式输入到计算机文件中，表 2.4 是一个根据抽样调查得来的数据生成的数据表。在该数据表中，每一列代表一个变量，如性别，每一行代表一个个体，这样的—个数据表通常叫做数据文件，利用 SPSS 进行数据分析的第一步就是建立这样的数据文件。

表 2.4 一次抽样调查的数据表

人员编号	性别	部门	体检日期	体重	健康状况
1	女	通信学院	08/10/2015	55	好
2	女	计算机学院	08/10/2015	46	好
3	女	外语学院	08/10/2015	50	一般
4	男	通信学院	08/10/2015	56	差
5	男	管理学院	08/11/2015	51	差
6	男	光电学院	08/11/2015	53	好
7	女	光电学院	08/11/2015	50	一般
8	男	通信学院	08/12/2015	50	好
9	女	计算机学院	08/12/2015	45	一般
10	男	管理学院	08/13/2015	56	好

SPSS 数据文件是一种有别于其他文件（如 Word 文档、文本文件）的特殊格式文件。它是一种有结构的数据文件，由数据的结构和内容两部分组成，其中数据的结构用来定义数据表中每一

列的属性,包括变量名、变量数据类型、标签、数据缺失情况等必要信息,数据的内容就是数据表中的每一行,即那些待分析的具体数据。

SPSS 数据文件与一般文本数据的不同在于,一般文本文件仅有纯数据部分,而没有关于结构的描述。正因如此,SPSS 数据文件不能像一般文本文件那样可以直接被大多数编辑软件读取,而只能在 SPSS 软件中打开。

基于上述特点,建立 SPSS 数据文件时应完成两项任务:

➤ 描述 SPSS 数据的结构。

➤ 录入、编辑 SPSS 数据。

为了能够顺利地使用 SPSS 处理数据,我们首先应熟悉如下概念:

(1) 个案。在数据处理中,一个研究对象就是一个个案,相当于一条记录,在数据表格中表现为“一行”。每一个个案记录的是一个研究对象各个属性的具体数值,如表 2.4 中人员编号 1 的信息,包括性别、体检日期、体重、健康状况等。

(2) 样本。样本是指具有共同属性的所有研究对象,如某学校一年级学生的所有信息。样本含多个个案,在数据表格里表现为“ n 行”,如表 2.4 中 10 个人的体检信息。

(3) 变量。SPSS 中的变量相当于数据库中的“字段”,在数据表格里表现为“一列”。如表 2.4 中,“人员编号”、“性别”、“部门”等都是变量名。

(4) 变量值。在 SPSS 系统里,单元格中的数值就是变量值。如表 2.4 中的第一个个案,变量“性别”的值为“女”。

2.2.3 输入数据建立数据文件

SPSS 数据的结构指对要分析的数据表中的每一列定义一个变量并描述其相关属性,变量的属性包括名称、类型、宽度、小数位数、标签、值、缺失值、列宽、对齐、测量、角色等信息。其中有些内容是必须定义的,有些是可以省略的。

1. 数据的结构定义

打开 SPSS 之后,进入数据编辑窗口,数据编辑窗口分为数据视图窗口和变量视图窗口。要建立新的 SPSS 数据文件首先需定义数据文件的结构,即定义新的变量,左键单击左下方的变量视图标签,得到如图 2-1 所示的变量定义窗口。

下面具体介绍变量定义窗口中各项的含义与设置。

(1) 名称

定义变量的名称。变量名是变量访问和分析的唯一标志,在定义 SPSS 数据文件的结构时,应首先给出每列变量的变量名。变量的命名规则一般如下:

➤ 每个变量名必须是唯一的,不允许重复。允许汉字作为变量名,汉字总数一般不超过 4 个。

➤ 变量名不能包含空格。


➤ 高版本 SPSS 的变量名长度多达 64 位,但是由于低版本 SPSS 变量名长度应在 8 位之内,为了避免与低版本及其他软件出现兼容问题,高版本变量名一般仍控制在 8 位之内且尽量避免使用中文,必要的中文说明可以放在“标签”栏中。

➤ 变量名不能与 SPSS 的保留字相同。SPSS 的保留字包括 all、by、eq、ge、gt、left、ne、not、or、to、with。系统不区分变量名的大小写。

➤ 应避免用句点结束变量名,因为句点可能被解释为命令终止符。只能使用命令语法创建以句点结束的变量,不能在创建新变量的对话框中创建以句点结束的变量。

总之，在为变量命名时，为方便记忆，变量名最好与其代表的含义相对应，做到见名知义。如果变量名不符合 SPSS 的命名规则，系统会自动给出错误提示，如果没有给变量命名，SPSS 会给出默认的变量名，以字母“VAR”开头，后面补足 5 位数字，如 VAR00001，VAR00012 等。

(2) 类型

选择变量类型。左键单击“类型”栏后的按钮，弹出如图 2-2 所示的对话框。SPSS 最基本的变量类型有数值型、日期型和字符串 3 种。每种类型都有默认的列宽度和小数位，通常数值型变量默认宽度为 8，小数位数为 2，字符型变量默认宽度为 8，这两种类型变量的默认宽度都可修改，而日期型变量则固定宽度为 10，不能修改。

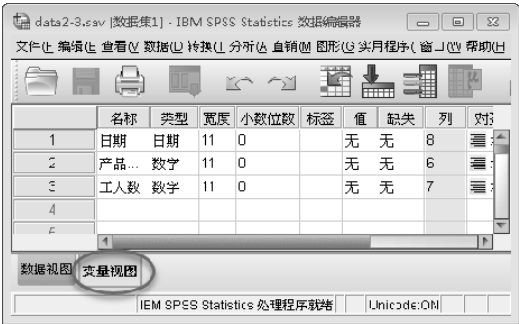


图 2-1 SPSS 变量定义窗口



图 2-2 “变量类型”对话框

根据基本的变量类型进一步细化，SPSS 的数据类型一共有 8 种，各种类型如表 2.5 所示。

表 2.5 SPSS 变量类型说明

中 文 名	说 明
标准数值型变量	值为数字的变量。值以标准数值格式显示，数据编辑器接受以标准格式或科学记数法表示的数值。默认总长度 8 位，小数位 2 位
带逗号的数值型变量	默认总长度 8 位，小数位 2 位，其值在显示时整数部分从右至左每 3 位用一个逗号做分隔符，值的小数指示符右侧不能包含逗号
带圆点的数值型变量	默认总长度 8 位，小数位 2 位，其值在显示时整数部分从右至左每 3 位用一个圆点做分隔符，值的小数指示符右侧不能包含句点
科学记数法数值型变量	默认总长度 8 位，小数位 2 位，它的值以嵌入的 E 及带符号的 10 次幂指数形式显示。数据编辑器为此类变量接受带或不带指数的数值。指数前面可以加上带符号（可选）的 E 或 D，或只加上符号，例如，123、1.23E2、1.23D2、1.23E+2 及 1.23+2
日期型变量	既可表示日期又可表示时间，用户可根据实际情况自行选择。其值以若干种日历—日期或时钟—时间格式中的一种显示，从列表中选择一种格式，输入日期时可以用斜杠、连字符、句号、逗号或空格作为分隔符
美元符号型数值变量	主要用来表示货币数据，显示时前面带美元符号（\$），每 3 位用逗号分隔，并用句点作为小数分隔符。可以输入带或不带前导美元符号的数值
定制货币型变量	一种数值变量，其值以定制货币格式中的一种显示，定制货币格式是在“选项”对话框的“货币”选项卡中定义的。定义的定制货币字符不能用于数据输入，但显示在数据编辑器中
字符串型变量	默认总长度 8 位，字符串值可以包含任何字符，可包含的最大字符数不超过定义的长度
受限数字	“数字”类型使用数位分组设置，而“受限数字”不使用数位分组

(3) 宽度

设置变量数字位数或字符个数。一般无须调整，直接采取默认值。它的大小可通过“宽度”栏右边的微调按钮调整，也可以通过图 2-2 的“宽度”选项进行调整。

(4) 小数位数

若变量类型为数值型，则可设置变量的小数位数，其他类型的变量则不能设置。小数位数默认为 2 位，也可在图 2-2 的“小数位数”框中输入数字进行小数位数的设置。

(5) 标签

定义变量名标签。考虑到与低版本的兼容问题，变量名最好限制在 8 位以内，并且尽量避免使用中文，这就有可能无法描述清楚变量的信息，此时就可在标签中对变量名做进一步的说明。利用“标签”栏，可以对变量详细说明，大大方便了用户对变量的理解。

(6) 值

这里的“值”指的是变量的值标签，值标签是对变量的可能取值附加的进一步说明，标签内容最多可以有 120 个字符，通常仅对分类变量的取值指定值标签。


对变量值附加标签值有重要的作用，例如我们定义一个变量“Departmt”，代表某所大学的学院或部门，准备将它作为分类变量参与数据文件的统计分析，可以将它定义为一个字符型变量，也可以定义为一个数值型变量。如果将它定义为一个字符型变量，则由于该校有众多的系和部门，在输入观测值时必须输入系或部门名称，这将大大地增加键盘输入的工作量。如果将它定义为一个数值型变量，日后在阅读数据文件的时候，常常又可能

不明确变量值的含义。如果将各系或部门的名称作为变量各个值的标签，在值标签开启的状态下，要输入各系或部门的名称，只需要输入对应的值，而在数据窗口变量值的单元格里却显示该变量值对应的值标签，既减轻了输入的工作量，又可以一目了然地了解变量值的意义。

例如，将变量 Departmt 定义为数值型变量时，按照表 2.6 所示定义值标签。

表 2.6 变量 Departmt 的值标签定义

变量值	变量值标签
1	通信学院
2	计算机学院
3	管理学院
4	光电学院
5	外语学院

左键单击图 2-1 中“值”一栏右边的按钮，弹出如图 2-3 所示的“值标签”对话框。在“值”栏中输入 1，“标签”栏中输入对应变量值的标签“通信学院”，当这两栏里输入了内容后，左边第一个按钮“添加”由灰色不可用变为可用，单击它可将输入的值标签添加到最下面的文本框中，如图 2-3 所示。用相同的方法，可添加其余的值标签。输入完所有的变量值标签后，单击“确定”按钮使对变量值标签的设置有效。如果输入有误，可单击文本框中显示的错误标签，然后单击“更改”按钮修改已经输入的标签，单击“删除”按钮可删除不需要的标签。

定义完变量值标签后，在 SPSS 主窗口的菜单栏中选择“查看→值标签”，如图 2-4 所示，“值标签”一项前的复选框被选中，则在 SPSS 主窗口中经过变量值标签定义的数值型变量显示为所定义的标签，例如，在“Departmt”一列显示的是文本“通信学院”、“计算机学院”等，而不是 1、2 这样的数值。

(7) 缺失

在统计分析的数据收集过程中，有时会因为某些原因产生所记录的数据失真，或者没有记录等异常情况。例如，在调查问卷中，被调查者没有填写调查表必须填写的某些数据，称为缺失值；学生的体检表中某学生的年龄为 60 岁，这显然是一个失真数据，不能使用，但其他数据在分析过程中还可以使用。这些情况称为数据缺失或数据不完全，在统计分析中这些数据是不能使用的。


SPSS 统计软件的另一特点就是可以通过制定缺失值的方式来定义缺失数据，这样就可以更好地利用其他的有效数据。在“缺失”栏单击按钮，弹出如图 2-5 所示的对话框。



图 2-3 “值标签”对话框



图 2-4 值标签显示设置

在“缺失值”对话框中包括 3 个单选按钮，其含义分别如下。

- 无缺失值：即对缺失值不做处理，不指定缺失值。
- 离散缺失值：对数值型或字符型变量，用户指定 1~3 个特定的离散值来代替缺失值。
- 范围加上一个可选的离散缺失值：选择该单选按钮，表示对数值型变量，用户缺失值定义在一个连续的闭区间以外的离散值，“下限”和“上限”分别表示连续区间的左右端点，在“离散值”文本框中输入区间以外的一个确定值。


(8) 列

定义变量在数据窗口中显示的宽度，列宽度只影响数据编辑器中的值显示，更改列宽不会改变变量已定义的宽度。

(9) 对齐

定义变量值显示的对齐方式，控制着数据视图中数据值或值标签的显示。默认对齐方式为数值变量在右边，字符串变量在左边，此设置只影响数据编辑器中的显示。

(10) 测量

在“测量”栏单击  按钮，弹出如图 2-6 所示的下拉列表，该菜单中列出了“标度”、“有序”、“名义”3 种标准的度量尺度供选择。该 3 种度量尺度与 2.2.1 节所介绍的统计数据度量尺度的概念有些差别，其对应关系如下：

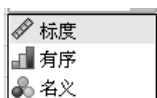


图 2-6 度量标准选项

- 标度——间隔尺度（定距尺度和定比尺度）。
- 有序——定序尺度。
- 名义——名义尺度。


变量属性“测量”中各选项所代表的图示均不相同，以数值类型的变量而言，度量标准设为“标度”，在变量清单中会以一支标尺表示（如图 2-6 中第一项前的标尺图标）；度量标准设为“有序”，在变量清单中会以长条图表示；度量标准设为“名义”，在变量清单中会以 3 个饼图表示。

在定义变量的“测量”属性时，究竟应该选择什么类型的度量尺度，应该视变量的数据类型和统计分析的需要而定，有序尺度和名义尺度可以是数值类型和字符串类型的变量，而标度尺度对应的变量类型只能是数值型；有些统计分析对于变量的标度尺度有一定的要求，尤其是名义尺度和标度尺度（间隔尺度），以独立样本 T 检验与方差分析而言，其自变量必须为名义尺度或有序尺度（定序尺度），而因变量必须为标度尺度。分析者若在变量的度量尺度属性的设置上界定清楚，则之后的统计分析会更为简便。



图 2-5 “缺失值”对话框

(11) 角色

在统计分析的某些对话框支持可用于预先选择分析变量的预定义角色，当打开其中一个对话框时，满足角色要求的变量将自动显示在目标列表中。在“角色”栏单击按钮，弹出如图 2-7 所示的下拉列表，该菜单中列出了“角色”属性中可以设置的选项。

各个选项的作用如下。

- 输入：变量将用作输入（例如预测变量、自变量）。
- 目标：变量将用作输出或目标（例如因变量）。
- 两者：变量将同时用作输入和输出。
- 无：变量没有角色分配。
- 分区：变量用于将数据划分为单独的训练、检验和验证样本。
- 拆分：设定此角色是为与 SPSS Modeler 相互兼容，具有此角色的变量不会在 SPSS Statistics 中用作拆分文件变量。



图 2-7 角色选项

在默认情况下，为所有变量分配“输入”角色。这包括外部文件格式的数据和 SPSS Statistics 18 之前版本的数据文件，角色分配只影响支持角色分配的对话框。

2. 数据的录入

(1) 录入数据的一般方法


定义了所有变量后，单击数据编辑窗口的“数据视图”标签，即可在数据视图中输入数据。数据编辑窗口中黑框所在的单元为当前的数据单元，表示用户正在对该数据单元录入数据或正在修改该单元中的数据。因此，在录入数据时，用户应首先选中要输入数据的单元格。

数据录入时可以逐行录入，即录入完一个数据后，按 Tab 键，焦点移动到本行的下一个变量列上。也可以逐列，也就是按照变量录入数据，录入完一个数据后回车，焦点移动到本列的下一行上。除了直接录入之外，SPSS 还可以直接复制粘贴 Excel 和 Word 表格中的数据，但要求数据表的表头与 SPSS 文件的结构相同。同时，SPSS 中的数据也可以直接粘贴到 Excel 和 Word 之中，这大大方便了用户对数据的编辑。

SPSS 数据录入有一项特殊功能，就是连续粘贴相同值。例如，需要连续录入相同变量值的时候，可以先录入一项，然后单击鼠标右键，在弹出的快捷菜单中选择“复制”，再拖动鼠标选中

所有要录入该值的单元格，单击鼠标右键，在弹出菜单中选择“粘贴”。这时，所有的单元格都已经同时粘贴上该值，而无须逐个粘贴了。

(2) 录入带有变量值标签的数据

输入定义了变量值标签的数据时，可以直接输入变量值，也可以单击要输入数值的单元格，出现，单击箭头按钮，展开一个列有该变量所有值标签的下拉式列表框，如图 2-8 所示，从中选择值标签即可。

No	sex	departmt	date
1	女	通信学院	08/10/2004
2	女	通信学院	08/10/2004
3	女	计算机学院	08/10/2004
4	男	管理学院	08/10/2004
5	男	光电学院	08/11/2004
6	男	外语学院	08/11/2004

图 2-8 通过变量值标签录入数据

☆说明☆

◆ 在图 2-8 中，“部门（departmt）”列显示的是“通信学院”、“计算机学院”这样的文本，但在数据文件中存储的是“1”、“2”这样的数值。怎样才能在数据编辑窗口显示“通信学院”这样的值标签，而不是存储的数值“1”呢？只需选择“查看”菜单中的“值标签”选项即可，如图 2-4 所示，系统默认的是显示变量值而不是变量值标签。

3. SPSS 数据文件建立实例

【例 2-1】 一次抽样调查的数据如表 2.4 所示，定义的各变量及其主要属性如表 2.7 所示。在定义 SPSS 数据的结构时，最常用的属性为变量名、数据类型、变量标签、变量值标签，表 2.7 给出了这几个属性的设置，其余属性可用默认属性，也可自行调整。（参见数据文件：data2-1.sav。）

表 2.7 各变量主要属性

变量名	数据类型	变量标签	变量值标签
No	数值类型	人员编号	无
Sex	字符串型	性别	0—女 1—男
Departmt	数值类型	部门	1—通信学院 2—计算机学院 3—管理学院 4—光电学院 5—外语学院
Date	日期型	体检日期	无
Weight	数值类型	体重	无
Health	数值类型	健康状况	1—差 2—一般 3—好

第 1 步 定义结构。

根据 2.2.3 节介绍的定义 SPSS 数据结构的方法，在变量定义窗口中建立数据文件中所涉及
的各变量及属性，图 2-9 是建立好的各变量及其属性。

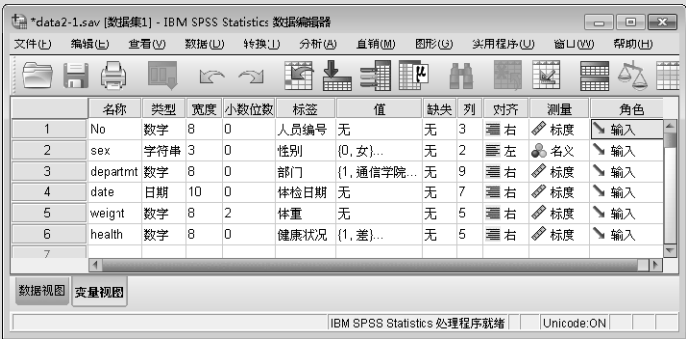


图 2-9 变量定义窗口中定义好的各变量及其属性

第 2 步 录入数据。

每定义一个变量后，按列方向录入数据，或定义完全部变量后，按行或者按列录入数据。对于定义了值标签的变量，如 Departmt，可以用数据录入方法中介绍的带有变量值标签的数据录入方法来录入。

2.2.4 从其他数据文件导入数据建立数据文件

以上几节介绍了建立 SPSS 数据文件的一般方法，在实际应用中，通常一批待分析的数据已经保存成了其他格式的数据文件，例如数据库格式的文件、Excel 文件等，若要用 SPSS 分析这些数据，就需要将这些数据转换到 SPSS 中，形成 SPSS 能处理的数据。因此，读取其他格式的文件并将其转换为 SPSS 格式的数据，是另一种建立 SPSS 数据文件的方法。从其他数据文件导

入数据的方式主要有：直接打开、用数据库查询方式打开、从文本文件导入几种方法。这几种方法中最简单的是直接打开，但有的数据文件不能直接打开，此时可以采用数据库查询的方式打开，而从文本文件导入数据则是一种针对纯文本数据文件的打开方式。下面分别介绍这 3 种导入数据的方式。

1. 直接打开

SPSS 可直接打开很多类型的数据文件，选择菜单“文件→打开→数据”，弹出“打开文件”对话框，左键单击“文件类型”，即可看到 SPSS 所能打开的数据文件类型，如表 2.8 所示。

表 2.8 SPSS 能直接打开的数据文件类型

文件扩展名	具体描述
SPSS (*.sav)	当前版本 SPSS 23 数据文件
SPSS/PC+ (*.sys)	低版本 SPSS 数据文件
Systat (*.syd *.sys)	Systat 格式数据文件
SPSS Portable (*.por)	SPSS 的 ASCII 数据文件
Excel (*.xls *.xlsx *.xlsm)	各种版本的 Excel 数据文件，此种数据格式常用
Lotus (*.w*)	Lotus 数据文件
Sylk (*.slk)	Sylk 数据文件
dBase (*.dbf)	dBase 数据文件，Foxpro 下的 dbf 文件需转换为 dBase 文件才能打开
SAS (*.sas7bat,*.sd7,*.sd2,*.ssd01,*.xpt)	各种版本和类型的 SAS 数据文件，一种统计学软件的数据文件格式
Stata (*.dta)	Stata v4-8
文本文件 (*.txt, *.dat, *.csv, *.tab)	以记事本格式保存的数据文件

SPSS 能直接打开的数据文件类型有很多，其中导入 Excel 类型的数据文件在实际操作中用得比较多，下面详细介绍读取 Excel 文件的过程。

图 2-10 是 Excel 格式的数据，现在要将其数据导入 SPSS 中进行分析，要求 Excel 数据文件中的第一行作为 SPSS 数据文件的变量名，以下各行是要分析的数据，如图 2-10 所示。

	A	B	C	D	E	F	G	H	I	J
1	实验准备	讲解示范	实验指导	教学方法	语言文字	教学手段	课堂管理		变量名称	
2	85	59	98	58	83	89	82			
3	65	40	73	16	45	70	50			
4	60	84	61	65	71	97	97			
5	31	49	43	57	45	62	87			
6	87	49	51	82	61	98	75			
7	59	85	42	44	66	34	72			
8	80	53	91	29	94	99	77			
9	57	44	50	72	82	11	47			
10	47	47	45	12	88	78	45			
11	61	51	74	34	88	71	51			
12	43	57	40	57	53	43	68			
13	91	63	64	85	47	54	74			
14	61	64	89	43	37	85	66			
15	92	93	57	31	85	57	64			
16	48	76	49	17	45	24	58			

图 2-10 Excel 数据文件

第 1 步 选择菜单“文件→打开→数据”，打开如图 2-11 所示的对话框，选择文件类型为 Excel，找到要读取的 Excel 文件，弹出如图 2-12 所示的对话框。

第 2 步 设置读取 Excel 文件工作表的范围，SPSS 会自动判断读取范围为工作表的所有数据，若只读取部分数据则更改读取范围，否则就用默认范围。如果要将 Excel 文件中第一行数据作为 SPSS 数据文件的变量名，则应勾选“从第一行数据中读取变量名”前的复选框，默认该选项已经勾选。设置好后，单击“确定”按钮完成 Excel 文件的读取，读取后的结果如图 2-13 所示。



图 2-11 “打开数据”对话框

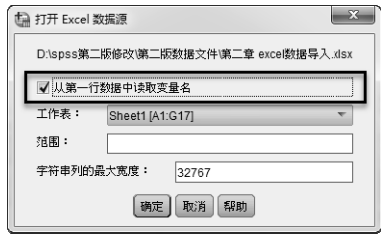


图 2-12 “打开 Excel 数据源”对话框

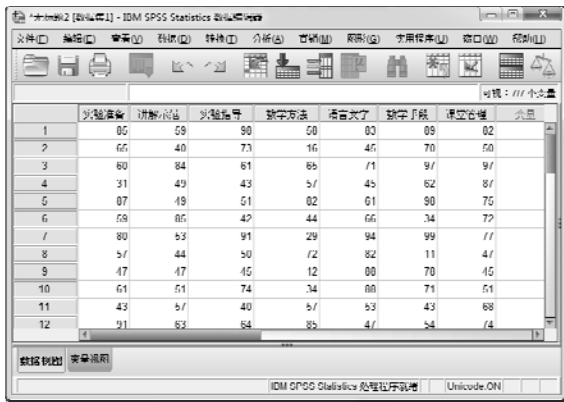


图 2-13 导入 Excel 数据后的结果

☆说明☆

- (1) 读取的 Excel 文件不能是打开状态，否则读取的时候会出错。
- (2) 如果 Excel 工作表文件第一行或指定读取区域内的第一行存储了变量名信息，则应选中“打开 Excel 数据源”对话框的复选框“从第一行数据中读取变量名”，即以工作表第一行或指定读取区域内第一行的文字信息作为 SPSS 的变量名，如果不选此项，SPSS 的变量名将自动取名为 V1、V2 等。

2. 数据库查询方式打开文件

如果数据为数据库格式的文件，可以用数据库查询的方式导入数据到 SPSS 中。其操作步骤如下：

第 1 步 选择菜单“文件→打开数据库→新建查询”，弹出如图 2-14 所示的“数据库向导”窗口，显示了所有可以打开的数据源类型。

第 2 步 用户根据打开文件的向导选择要打开的文件类型并逐步打开文件。

其实从前面的讲解可以发现，直接打开方式已经可以打开很多常见类型的数据文件了，但是当与 SQL Server、Oracle 等大型数据库进行数据交换时，直接打开数据文件往往不行，所以此时要使用数据库查询的方式打开数据文件。另外，如果用户使用的 SPSS 版本不稳定，那么对于简单的 Excel 文件有时也无法直接打开，此时也可以通过数据库查询的方式打开。



图 2-14 “数据库向导”窗口

3. 从文本文件导入数据

文本格式的数据文件是一种最通用的数据文件，SPSS 提供了专门读取文本文件的功能。选择菜单“文件→读取文本数据...”，弹出“打开文件”对话框，选择要导入的文本文件名后会出现文本数据导入的向导，该向导是一个分为 6 步的打开向导，只需按照向导一步一步地设置，即可读取文本文件的数据到 SPSS 中，限于篇幅，本书不再赘述，详细过程可参见其他书籍。

2.3 数据文件的编辑

2.3.1 数据文件的合并

数据文件的合并是把外部数据与当前数据合并成一个新的数据文件，SPSS 提供两种形式的合并：一是横向合并，指从外部数据文件中增加变量到当前数据文件中；二是纵向合并，指从外部数据文件中增加观测数据到当前数据文件中。

1. 横向合并

横向合并的效果如图 2-15 所示，横向合并有两种方式：一是从外部数据文件中获取一些变量数据，加入当前数据文件中；二是按关键变量合并，要求两个数据文件有一个共同的关键变量，而且两个数据文件的关键变量之中还有一定数量相同值的观测量。

横向合并的具体步骤如下：

第 1 步 打开“合并文件”对话框。

在当前数据文件中选择菜单“数据→合并文件→添加变量”，弹出如图 2-16 所示的对话框。

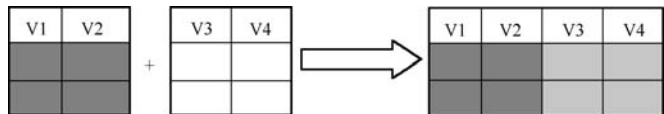


图 2-15 添加变量

第 2 步 打开“添加变量”对话框。

选择“外部 SPSS Statistics 数据文件”单选按钮，单击“浏览...”按钮选择需要合并的 SPSS 数据文件后，单击“继续”按钮，打开如图 2-17 所示的“添加变量”对话框。

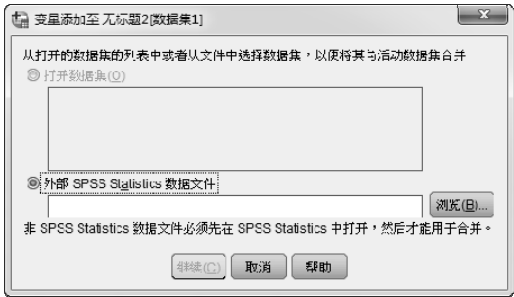


图 2-16 “合并文件”对话框

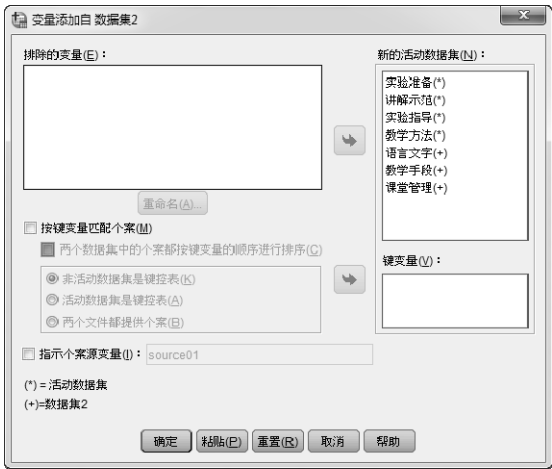


图 2-17 “添加变量”对话框

第 3 步 选择需添加的变量。

在“添加变量”对话框中，选择需要添加的变量到“新的活动数据集”框中，单击“确定”按钮即可完成合并操作，并在当前数据编辑窗口显示合并后的数据文件。

☆说明☆

- (1) 变量名旁标有“*”的为当前工作数据文件中的变量，标有“+”的为外部数据文件的变量。
- (2) 如果要将“排除的变量”列表框中的同名变量加入合并变量的数据文件中，可以单击“重命名”按钮，对变量重命名后再将此变量加入到“新的活动数据集”列表框中。
- (3) 如果按照关键变量合并，则需选择合并的关键变量，并且两个数据文件先按关键变量以相同的方式排序，但是，当两个数据文件具有相同的个案数并且排列顺序一致时，则不需要指定关键变量，只需要单击“确定”按钮即可。

2. 纵向合并

纵向合并即增加个案，合并前后的变化如图 2-18 所示，是在两个具有相同变量的数据文件中，将其中一个数据文件的个案追加到当前数据文件的个案中，形成新的数据文件。

纵向合并数据文件的操作方法同横向合并的方法类似，操作步骤不再赘述，但需注意以下几点：

- 两个待合并的 SPSS 数据文件的内容合并起来应具有实际意义。
- 两个数据文件的结构最好一致。
- 不同数据文件中含义相同的变量最好用相同的变量名，数据类型要相同。

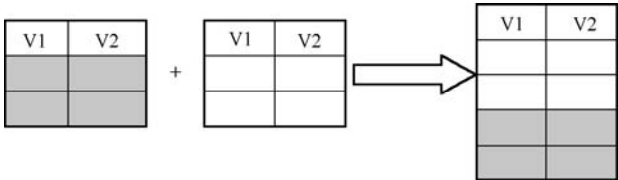


图 2-18 纵向合并

2.3.2 数据文件的拆分

对数据文件进行拆分，SPSS 23 版本在“数据”菜单中提供了两种方法：一是“拆分文件...”，这种拆分并不是要把一个数据文件分成几个数据文件，而是按照需求，根据变量对数据进行分组，为以后的分组统计分析提供便利；二是“拆分为文件”，这种拆分是将拆分后的数据写入新的 SAV 文件，按拆分变量的值或值标签生成多个 SAV 文件。

1. 拆分文件

在进行数据分析的时候，有时需要对数据文件按某个变量进行拆分，这种拆分并不是要把数据文件分成几个，而是根据实际情况，根据变量对数据进行分组，为以后的分组统计分析提供便利。

【例 2-2】 表 2.9 中是各产品销售的数据，分别统计各产品的销售总量及销售总额。（参见数据文件：data2-2.sav。）

表 2.9 销售员销售产品的统计表

姓名	日期	产品	数量	单价	金额
李汉青	2010-1-1	彩电	42	3200	134400
张三中	2010-1-2	彩电	40	3200	128000
李开	2010-1-3	空调	3	3200	9600
张国华	2010-1-4	微波炉	24	2100	50400
王三	2010-1-5	热水器	24	2300	55200
刘利国	2010-1-6	彩电	12	3200	38400
杜为	2010-1-7	洗衣机	5	2200	11000
吴兵	2010-1-8	洗衣机	48	2200	105600
张国华	2010-1-4	微波炉	1	2100	2100
王三	2010-1-5	热水器	11	2300	25300
刘利国	2010-1-6	彩电	50	3200	160000

所有产品的数据都在一个数据文件中，此时需要在分析之前依据“产品”这个间断变量的水平，将数据文件进行拆分，这里的拆分，并非将一个数据文件拆分为两个或若干个独立的数据文件，而是按产品变量进行分组，分别求出不同产品的描述性统计量，拆分后进行描述统计的效果如图 2-19 所示。

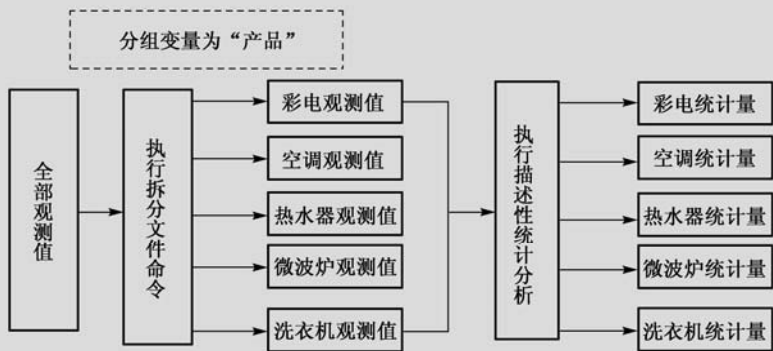


图 2-19 数据文件拆分后的描述统计效果

在 SPSS 中进行数据拆分后统计产品的销售总量和销售金额的步骤如下:

第1步 数据组织。

首先将表 2.9 中的数据整理成 SPSS 中的数据文件，建立“姓名”、“日期”、“产品”、“数量”、“单价”和“金额”6 个变量，保存为 data2-2.sav 文件。

第2步 打开主对话框。

选择菜单“数据→拆分文件...”，弹出如图 2-20 所示的“拆分文件”对话框。

第3步 选择拆分方式。

按照产品类型拆分数据,选择“比较组”,激活“分组方式”栏,选中“产品”变量移入其中,单击“确定”按钮结束。

拆分后的数据文件将按“产品”变量排序，并将显示在数据编辑窗口中代替原文件。数据拆分后将对后面的统计分析一直起作用，即无论进行哪种统计分析，都将按拆分变量分组进行不同组别的分析计算，例如，对表 2.9 的数据按“产品”变量拆分成分组后，统计销售数量的总和时将按照产品分组进行总计。

第4步 按产品分组统计销售总量和销售总额。

选择菜单“分析→描述统计→描述...”，弹出如图 2-21 所示的“描述”对话框，选择变量“金额”、“数量”进行分析，单击“选项”按钮设置要计算的统计量，此处只需统计金额和数量的和，设置好后单击“确定”按钮，得到如表 2.10 所示的统计结果。

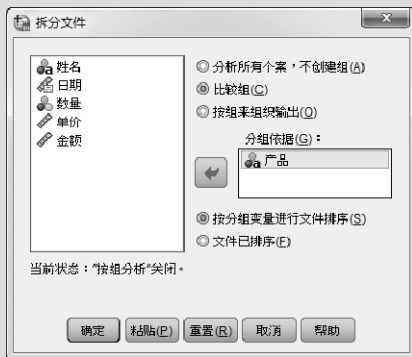


图 2-20 “拆分文件”对话框

表 2.10 描述统计量

产品		个案数	总和
彩电	金额	4	460800
	数量	4	144
	有效个案数（成列）	4	
空调	金额	1	9600
	数量	1	3
	有效个案数（成列）	1	
热水器	金额	2	80500
	数量	2	35
	有效个案数（成列）	2	
微波炉	金额	2	52500
	数量	2	25
	有效个案数（成列）	2	
洗衣机	金额	2	116600
	数量	2	53
	有效个案数（成列）	2	



图 2-21 “描述”对话框

☆说明☆

(1) “拆分文件”对话框中，“比较组”与“按组组织输出”的区别在于：前者将分组统计结果输出在同一张表格中，以便于不同组之间的比较；后者将分组统计结果分别输出在不同的表格中。通常选择第一种输出方式。

- (2) 若要取消数据拆分, 只需选择“分析所有个案, 不创建组”即可。
- (3) 对数据可以进行多重拆分, 类似于数据的多重排序, 多重拆分的次序决定于选择拆分变量的前后次序。

2. 拆分为文件

拆分为文件是将数据文件按拆分变量的值或值标签, 拆分为多个数据文件, 例如例 2-2 中的数据文件, 若将“产品”作为拆分变量, 由于“产品”变量有彩电、空调等 5 个值, 因此将拆分为 5 个数据文件, 具体操作步骤如下:

第 1 步 数据组织。

打开在例 2-2 中整理好的数据文件 data2-2.sav。

第 2 步 打开“将数据集拆分为单独的文件”对话框。

选择菜单“数据→拆分为文件”, 弹出如图 2-22 所示的“将数据集拆分为单独的文件”对话框。



图 2-22 “将数据集拆分为单独的文件”对话框

第 3 步 拆分文件生成设置。

首先, 选择拆分变量, 将“产品”变量移入右面的“按以下变量拆分个案”框中, 则拆分文件按“产品”变量的值对个案进行分组, 并将各个分组拆分生成不同的文件; 其次, 设置输出文件目录和输出文件名, 设置文件名的方法有两种:

- (1) 以拆分变量的值作为输出文件名, 如图 2-23 所示, 只需设置输出文件目录, 输出文件名不用设置, 默认输出文件名以拆分变量的值作为文件名, 拆分后的效果如图 2-24 所示;
- (2) 以固定的文件名作为输出文件, 由于拆分后的文件名一样, 因此需要设置将文件放在不同的目录下, 要生成不同目录, 需要引用拆分变量, 在输出文件目录中以 `${变量名}` 格式进行引用, 其设置方法如图 2-25 所示, 输出文件目录设置为 `d:\target\${产品}`, 拆分时, 会在 d 盘生成目录 target, 并在 target 下以拆分变量的值生成目录, 拆分后的文件会放到对应的目录下, 其拆分后的效果如图 2-26 所示。

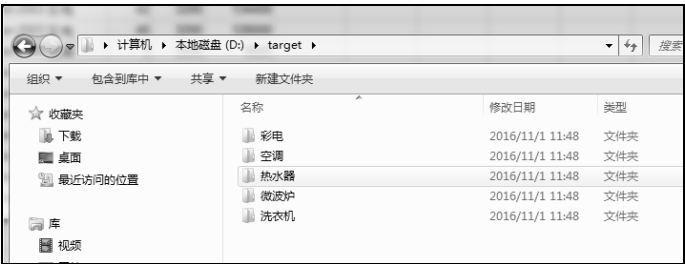


图 2-26 以固定文件名输出的结果

2.3.3 数据的选取

有时为了进行特定的分析，需要从所有的数据资料中选择部分数据进行统计分析。例如，对表 2.9 中的数据，只选取“彩电”这种产品进行分析，具体操作步骤如下：

第 1 步 数据组织。

打开例 2-2 中整理好的数据文件 data2-2.sav。

第 2 步 打开“选择观测量”对话框。

选择菜单“数据→选择个案...”，弹出如图 2-27 所示的“选择个案”对话框。

第 3 步 指定选择个案的方式。

选择“如果条件满足”选项，表示按指定条件选择观测量。

系统提供了几种选择观测量的方法，除了以上选择的按指定条件设置外，还有以下几种。

(1) 所有个案：所有的个案都选择。该选项可用于解除原来的个案选择。

(2) 随机个案样本：对个案进行随机抽样，即对数据编辑窗口中的所有个案进行随机筛选。包括两种方式的随机筛选：一是近似抽样，即输入抽样比例后由系统随机抽取；二是精确抽样，即要求从第几个观测量起抽取多少个。

(3) 基于时间或个案范围：顺序抽样，单击“范围...”按钮可以定义从第几个观测量到第几个观测量。

(4) 使用过滤变量：用指定的变量（只能为数字型变量）进行过滤，即依据过滤变量的取值进行样本选取，变量值为非 0 或非系统缺失值的个案将被选中。这种方法通常用于排除包含系统缺失值的个案。

第 4 步 设置选中个案的输出形式。

“过滤掉未选定的个案”是默认设置，通常选择此默认设置。

各输出形式的含义如下。

(1) 过滤掉未选定的个案：该输出形式在当前数据文件中自动生成一个名为 filter_\$ 的新变量，取值为 0 或 1，1 表示本个案被选中，0 表示未被选中，并在未选中的个案前做删除的标记。



图 2-27 “选择个案”对话框

(2) 将选定个案复制到新数据集：将选中的个案输出到新的数据文件中，设置新数据文件的文件名即可。

(3) 删除未选定个案：在当前数据文件中删除未选中的个案。

第5步 设置选择个案的条件。

单击“如果...”按钮，弹出如图 2-28 所示的“选择个案: If”对话框，从左侧的变量列表框中选择变量“产品”进入右面的框中，设置条件：产品="彩电"，（注意：在输入双引号时须在英文状态下输入。）个案选择结果如图 2-29 所示。



图 2-28 “选择个案: If”对话框



图 2-29 个案选择结果

经过以上步骤的操作后，以后的统计分析只会针对产品是“彩电”的个案，若要取消以上的个案选择，只需打开“观测量选择”对话框，选择其中的“全部个案”即可。

2.3.4 数据的加权

权重是统计学里的重要概念之一。在记录有大量数据的文件里，可能多次测量到同一观测量值，所谓权重是指同一个观测量值在所有的观测量里出现的次数或频率。SPSS 的观测量加权功能是在数据文件中选择一个变量，这个变量里的值是相应的观测量出现的次数，这个变量叫做权重变量，经过加权的数据文件叫做加权文件。

表 2.11 工人生产情况统计表

日 期	产 品 数 量	工 人 数
2004-1-1	20	3
2004-1-1	25	5
2004-1-1	30	3
2004-1-1	23	4
2004-1-1	20	4

【例 2-3】某工厂统计工人每天生产产品的数据，记录的结果中有些工人生产的产品个数是相同的，例如，生产产品 20 个的工人有 3 位，观测到的数据整理后形成表 2.11 中的数据，试统计产品数量的总和。（参见数据文件：data2-3.sav。）

要统计产品数量总和，则需要将变量“工人数”指定为权变量。将表 2.11 中的数据整理成 SPSS 数据文件 data2-3.sav，对观测量加权的具体

步骤如下。

第 1 步 数据组织。

不同的数据组织，会导致不同的分析方法和步骤，在本例中，可以有以下两种数据组织方法。

① 数据文件中只建立一个“产品数量”变量。

录入数据时应注意，不应只有表 2.11 中的 5 行数据，个案数应是所有的“工人数”相加，即 19 个个案，录入的数据应该有 19 行：3 行“产品数量”值为 20 的个案，5 行“产品数量”值为 25 的个案，以此类推。

这种数据组织方式不用加权处理，直接用“分析”菜单下的“描述统计”功能统计产品的数量总和即可。

② 数据文件中建立两个变量：“产品数量”和“工人数”

录入数据时按表 2.11 所列数据录入即可，但在进行统计计算前，必须进行加权处理，将变量“工人数”作为权重变量处理。

本例按第二种数据组织方式建立数据文件，保存为 data2-3.sav。

第 2 步 打开“个案加权”对话框。

选择“数据→个案加权”，弹出如图 2-30 所示的“个案加权”对话框。

第 3 步 设置加权变量。

选择“个案加权”，激活“频率变量”矩形框，从源变量列表中选择“工人数”变量移入此框中。

第 4 步 统计产品数量总和。

选择“分析→描述统计→描述”，进行产品数量总和的统计，统计结果如表 2.12 所示。

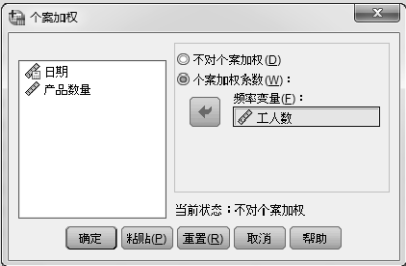


图 2-30 “个案加权”对话框

表 2.12 描述统计量

	N	和	均值
产品数量	19	447	23.53
有效的 N (列表状态)	19		

加权数据文件和未经加权的数据文件从数据窗口来看没有任何区别，它们的差异只有在调用统计分析过程后才能显现出来。data2-3.sav 数据文件加权前，选择“分析→描述统计→描述”计算产品的数量总和为 118，加权以后计算的产品数量总和为 447，显然加权以后的计算结果才是正确的。

从加权的含义不难理解, SPSS 中指定加权变量的本质是数据复制, 例如表 2.11 中的数据, 指定变量“工人数”为加权变量后, SPSS 将第一行的数据复制 3 行, 将第二行的数据复制 5 行, 等等。通过这样的处理, 可以达到将数据编辑窗口中的汇总数据还原为原始数据的目的。

☆说明☆

- (1) 一旦指定了加权变量, 在以后的分析处理中加权便一直有效, 直到取消加权为止。
- (2) 只有数值型的变量才能作为加权变量。

2.4 SPSS 数据加工

2.4.1 变量的计算

【例 2-4】图 2-31 是某高校学生对教师的各项教学指标评价的数据, 各项指标占总分的百分比分别是: 实验准备 (15%)、讲解示范 (15%)、实验指导 (20%)、教学方法 (15%)、语言文字 (5%)、教学手段 (10%)、课堂管理 (20%)。数据已经整理形成 SPSS 数据文件, 存储在 data2-4.sav 中, 利用 SPSS 提供的变量计算功能, 根据各指标计算出对教师评价的总分。(参见数据文件: data2-4.sav。)

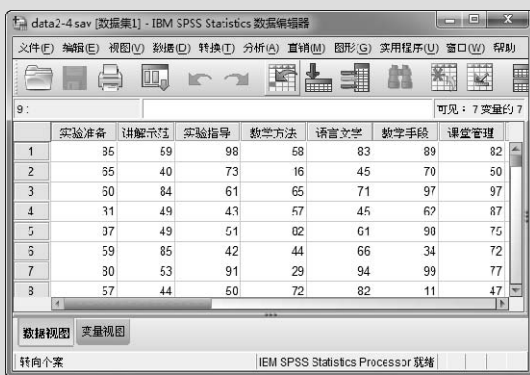
详细操作步骤如下。

第 1 步 数据组织。

数据文件中变量的建立如图 2-31 所示, 录入数据后保存为数据文件 data2-4.sav。

第 2 步 打开“计算变量”窗口。

选择菜单“转换→计算变量”, 弹出如图 2-32 所示的“计算变量”窗口。



	实验准备	讲解示范	实验指导	教学方法	语言文字	教学手段	课堂管理
1	35	59	98	58	83	89	82
2	35	40	73	16	45	70	50
3	50	84	61	65	71	97	97
4	31	49	43	57	45	67	87
5	37	49	51	02	61	90	75
6	59	85	42	44	66	34	72
7	30	53	91	29	94	99	77
8	57	44	50	72	82	11	47

图 2-31 教学指标评价数据



图 2-32 “计算变量”窗口

第 3 步 选择目标变量。

在“目标变量”框中输入目标变量名“总分”, 即存储计算结果的变量。

第 4 步 输入计算表达式。

从右边的变量列表窗口中选择用于计算的变量并加入“数学表达式”框中, 乘以相应的系数即可。

计算评价结果的表达式： $\text{实验准备} \times 0.15 + \text{讲解示范} \times 0.15 + \text{实验指导} \times 0.2 + \text{教学方法} \times 0.15 + \text{语言文字} \times 0.05 + \text{教学手段} \times 0.1 + \text{课堂管理} \times 0.2$ 。计算后的结果如图 2-33 所示，新增加一个“总分”变量，变量值为教师所得综合评价分。

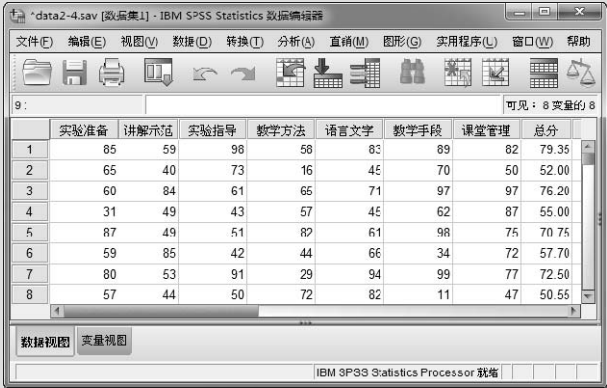


图 2-33 变量计算后的结果

以上操作过程中未涉及变量计算中提供的其他功能，在图 2-32 所示的“计算变量”窗口中还可以进行以下设置。

（1）“如果...”按钮

在进行变量计算时，有时要求对只满足条件的个案才计算。例如，在记录某班学生成绩的数据文件中，要了解女同学的学习情况，计算女同学的功课总分，可单击“如果...”按钮，在对话框中设置条件，条件设置的方法同个案选择的条件设置类似。

（2）“函数组”列表框

“函数组”列表框里列举了 SPSS 的所有函数组，单击任意一个组名，这一组中所有的函数和特殊变量将出现在“函数和特殊变量”框内，再单击任意一个函数名，这个函数的信息就出现在小计算器面板下的空白框中，供用户查询该函数的意义和用法。利用这些函数可以生成指定分布的随机数、给定参数的概率密度函数等。关于这些函数的具体使用方法，读者可以参考其他相关书籍，此处不再赘述。

2.4.2 数据可视分箱

SPSS 提供的数据可视分箱功能可将连续的数值型数据按由小至大的顺序加以分组（测量值由最低分至最高分分组），从而可将等距或比率变量转换为间断变量，其功能在于将连续数值数据分割为不同区段，区段的编码中最低分至第一个临界值的水平数值为 1（第一个区段），第二个区段的水平数值为 2，第三个区段的水平数值为 3，等等，第一个区段的水平数值一定是测量值中最低数值的那个区段，其水平数值内定为 1。

【例 2-5】 将图 2-33 数据中的“总分”变量按表 2.13 的标准分为 5 组，当评价结果小于 60 时，对应新变量的值为 1，变量值标签为“不合格”，以此类推。（参见数据文件：data2-5.sav。）

完成以上标准的变量组段划分，也就是将“总分”变量的值，离散化成不同的等级，详细操作步骤如下。

第 1 步 数据组织。

在例 2-4 中，生成了“总分”变量和其相应的值，存储为数据文件 data2-5.sav，本例中打开文件 data2-5.sav，进行变量的分箱处理。

表 2.13 “评价结果”变量分组的标准

“评价结果”变量的值	新变量的值	新变量值标签
<60	1	不合格
60~70	2	合格
70~80	3	中
80~90	4	良
90~100	5	优

第 2 步 打开“可视分箱”的主对话框。

选择“转换→可视分箱...”，弹出如图 2-34 所示的变量选择对话框，将需要进行变量值分组的变量“总分”选入“要分箱的变量”框，单击“继续”按钮进入如图 2-35 所示的“可视分箱”对话框。

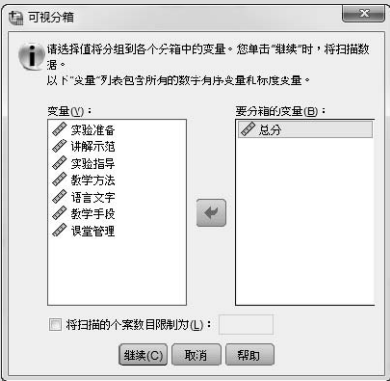


图 2-34 可视分箱的变量选择对话框



图 2-35 “可视分箱”对话框

在如图 2-34 的对话框下面的选项“将扫描的个案数目限制为”可设置参与分析的记录数目，指定数目放在后面的空白栏内，此选项适用于数据量很大时。

第 3 步 新变量名、变量标签设置和分割点的设置。

- (1) 在图 2-35 的“分箱化变量”框中输入新变量名“等级”，变量标签可设置也可不设置。
- (2) 在图 2-35 的“上端点”框中选择“排除(<)”，表示将已确定的分组断点的上限值归入下一个分组中，例如，总分小于 60 则评价等级为不合格，等于 60 则划入下一个组，即合格那一组。

(3) 单击“生成分割点...”按钮，弹出如图 2-36 所示的“生成分割点”对话框。选择“等宽区间”选项，即按照变量值等间距划分，在“第一个分割点位置”栏中输入第一个断点处的取值 60.00，也就是将最小值到 60 之间的数作为第一个分组组段。在“宽度”栏内输入一个组段内变量值的长度 10，再单击“分隔点数”一栏，系统会根据当前总分的值计算出分隔点的数量 4。单击“应用”按钮返回到图 2-35 所示的对话框。

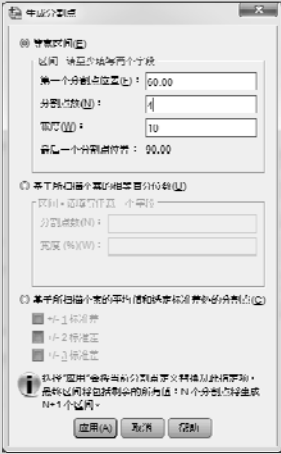


图 2-36 “生成分割点”对话框

(4) 经过步骤 (3) 的设置，回到图 2-35 所示的对话框时，在“值”一栏将出现各断点处的值，在“标签”一栏内可设置变量的值标签，如图 2-37 所示。

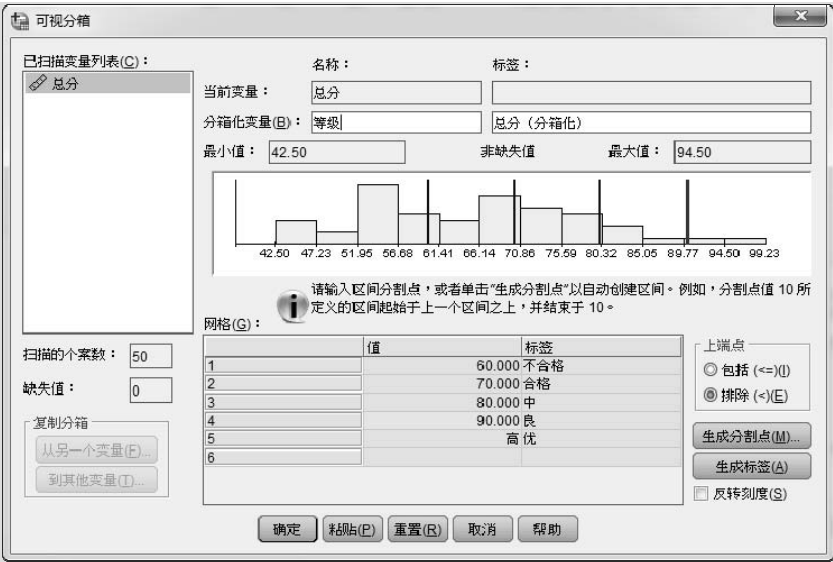


图 2-37 分割点设置好后的界面

第 4 步 完成分组设置。

单击图 2-37 中的“确定”按钮，提示“将创建一个新的变量”，确定以后在数据窗口创建一个变量“等级”，其结果如图 2-38 所示，图中新创建变量“等级”的值显示的是值标签。

实验准备	讲解示范	实验指导	教学方法	语言文字	教学手段	课堂管理	总分	等级
85	59	98	58	83	89	82	79.35	中
65	40	73	16	45	70	50	52.00	不合格
60	84	61	65	71	97	97	76.20	中
31	49	43	57	45	62	87	55.00	不合格
87	49	51	82	61	98	75	70.75	中
59	85	42	44	66	34	72	57.70	不合格
80	53	91	29	94	99	77	72.50	中
57	44	50	72	82	11	47	50.55	不合格
47	47	45	12	88	78	45	46.10	不合格
61	51	74	34	88	71	51	58.40	不合格
43	57	40	57	53	43	68	52.10	不合格
91	63	64	85	47	54	74	71.20	中
61	64	89	43	37	85	66	66.55	合格

图 2-38 创建分组变量以后的数据

除了以上的等间距划分外，在图 2-36 中还有另外两种分组方法。

(1) 基于所扫描个案的相等百分位数

按相等比例的观测值数目进行分组划分，在“分割点数”栏内输入断点的数目，系统自动将每组观测值数目的比例输出到“宽度”栏内。

(2) 基于所扫描个案的平均值和选定标准差处的分割点

基于变量的均值和标准差来产生组段划分。这一选项下 3 个复选框，分别指将断点设在以均值为中心、以+/-1、+/-2、+/-3 为标准差的断点。无论是否选择 3 个复选框，系统都将只产生一个断点，就是变量值的均值点。

2.4.3 数据重新编码

数据的重新编码是指给每个变量的观测值重新赋予一个新的值来描述它们的属性，并把相同的值分为一组，所以也称为变量的分组。例如，在例 2-5 中，将考评得分分为不同的等级，60 分以下等级为 1，60~69 分为 2，70~79 分为 3，80~89 分为 4，90 分以上为 5，将连续变量“总分”转换成 5 个相对应的等级，转换的方法是利用 SPSS 中提供的“可视分箱”功能，但“可视分箱”提供的功能只能将最低分至第一个临界值的水平数值赋值为 1（第一个区段），也就是只能将 60 分以下的等级赋值为 1，如果要组段重新划分为 90 分以上为 1，80~89 分为 2，70~79 分为 3，60~69 分为 4，60 分以下为 5，用“可视分箱”功能就不能实现此种重新编码，这时可以使用 SPSS 提供的另一个功能“重新编码”来实现。在“转换”菜单中提供了两个选项——“重新编码为相同的变量”、“重新编码为不同变量”，这两个选项实现数据重新编码的功能相同，差别在于：重新编码成相同变量时，新编码后数据会取代原先变量中的原始数据；重新编码成不同变量则会保留原始变量内的数据，新编码后的数据会新增一个变量名称。

【例 2-6】 将例 2-5 中数据的“总分”变量按照如下标准分为 5 组：

90 分以上为 1，80~89 分为 2，70~79 分为 3，60~69 分为 4，60 分以下为 5。用“转换”菜单中的“重新编码为不同变量”功能实现。（参见数据文件：data2-5.sav。）

详细操作步骤如下：

第 1 步 打开“重新编码为不同变量”对话框。

选择“转换→重新编码为不同变量”，弹出如图 2-39 所示的对话框。在“输出变量”框中输入新变量的名称“新的编码”，单击“变化量”按钮，“总分→？”将会变为“总分→新的编码”，重新编码后的值会写入新的变量“新的编码”中。



图 2-39 “重新编码为不同变量”对话框

第 2 步 设置编码转换规则。

单击图 2-39 对话框上的“旧值和新值”按钮，弹出如图 2-40 所示的对话框。在对话框中旧值的设置有 7 项选择，新值的设置有 3 项，根据转换规则，选择旧值的范围，再设置相对应的新值，单击“添加”按钮添加到“旧→新”列表框中，有几条转换的规则就应添加几次，设置好的转换规则如图 2-40 所示。单击“继续”按钮返回到图 2-39 的对话框中，单击“确定”按钮，生成如图 2-41 所示的重新编码结果。



图 2-40 旧值转换为新值的设置

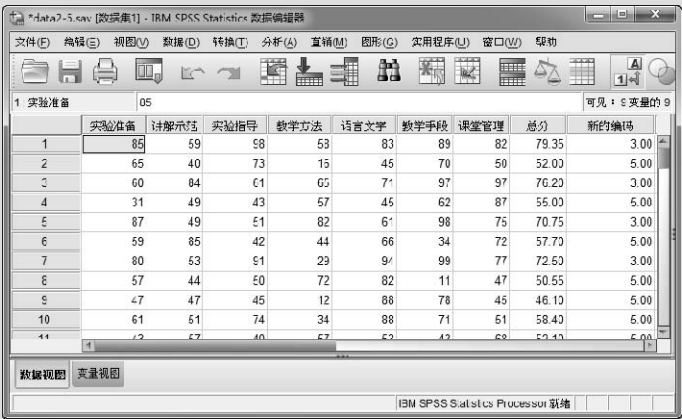


图 2-41 重新编码结果

2.5 思考与练习

1. 在 SPSS 中如何定义变量？变量名的命名有哪些要求？
2. 变量标签和变量值标签有什么区别？分别有什么作用？
3. 设计一份合格的调查问卷时，需要注意哪些问题？
4. 试设计一份关于大学本科生手机需求情况的调查问卷。要求格式正确，题目中要包含开放型和封闭型问题。
5. 以下是问卷调查表中的 3 个问题，调查结果如表 2.14 所示，请根据该调查表建立 SPSS 数据文件，并录入问卷调查结果，要求建立变量值标签。
 - (1) 您的性别是：男.....1 女.....2
 - (2) 您的家庭月收入大约是：（请包括所有工资、奖金、津贴等在内，以元为单位）单选
500~1000 元.....1 1000~1999 元.....2 2000~2999 元.....3 3000~3999 元.....4
4000~4999 元.....5 5000~5999 元.....6 6000~6999 元.....7 7000~7999 元.....8
8000~8999 元.....9 9000~9999 元.....10 10000 元及以上...11
 - (3) 您的教育程度：（指您受过的最高或正在接受的教育程度）单选

没有受过正式教育/小学.....1	初中.....2	高中/中专/技校.....3
大专/大学非本科/高职高专.....4	大学本科.....5	研究生及以上.....6

表 2.14 调查结果

问题一	2	2	2	2	1	1	1	2	2	2	1	1	2	2	2	2	1	1	1	2	2	1	1	2	1
问题二	3	1	2	1	1	5	5	5	5	4	3	7	4	5	3	4	6	6	2	10	4	11	4	3	3
问题三	5	5	5	5	5	5	5	4	5	4	5	5	2	2	3	4	3	5	5	5	3	3	5	5	3

6. 表 2.15 是某次调查居民家庭月收入的数据，试建立 SPSS 数据文件，并利用 SPSS 提供的“转换”菜单下的“重新编码为不同变量”功能将“家庭月收入”数据进行分组，新增一个分组变量，分组标准：2000~4000 元.....4，4000~6000 元.....3，6000~8000 元.....2，8000 元以上.....1。

7. 表 2.16 的数据是另外调查所得家庭月收入，是 Excel 格式的数据，将其导入到 SPSS 中并与第 6 题的数据合并，用 SPSS 提供的“转换”菜单下的“可视离散化”功能对“家庭月收入”进行重新分组，分组标准：2000 元以下为 1，2000~4000 元为 2，.....，以此类推，组距为 2000。试比较“重新编码为其他变量”功能和“可视离散化”功能有何不同。（参见数据文件：data2-6.xls。）

表 2.15 家庭月收入

编 号	家庭月收入（元）	编 号	家庭月收入（元）
1	3117	9	7149
2	2121	10	10336
3	3336	11	9432
4	2350	12	9322
5	5778	13	8739
6	11927	14	4351
7	6562	15	8205
8	4500	16	5960

表 2.16 家庭月收入

编 号	家庭月收入（元）	编 号	家庭月收入（元）
17	1900	25	8249
18	2500	26	21336
19	3650	27	8432
20	2350	28	7322
21	5978	29	5739
22	12927	30	9351
23	7562	31	9205
24	4780	32	1960

8. 表 2.17 是 2001 年华北 5 省市工业品产量，试用拆分数据文件的方式统计各省市工业品产量的总和。（参见数据文件：data2-7.sav。）

表 2.17 2001 年华北 5 省市工业品产量

序 号	省 市	工 业 品	产 量	序 号	省 市	工 业 品	产 量
1	北京	生铁	783.59	11	河北	水泥	4878.03
2	北京	钢	825.11	12	河北	塑料	40.4
3	北京	水泥	809	13	山西	生铁	2088.54
4	北京	塑料	75.6	14	山西	钢	606.77
5	天津	生铁	228.74	15	山西	水泥	1573.01
6	天津	钢	395.73	16	山西	塑料	3.2
7	天津	水泥	338.99	17	内蒙古	生铁	476.06
8	天津	塑料	73.6	18	内蒙古	钢	453.75
9	河北	生铁	21712.09	19	内蒙古	水泥	698.12
10	河北	钢	1969.65	20	内蒙古	塑料	5.8

9. 将文本文件“data2-8.txt”导入 SPSS，定义变量属性，然后任选分类变量进行个案排序并练习行列互换。

10. 将 Excel 数据文件“data2-9.xls”导入 SPSS，定义变量属性，然后选择分组变量练习文件拆分。

第 3 章 描述性统计分析

前面章节都是在为统计分析做准备，从本章开始，我们将正式进入统计分析的学习。数据处理和统计分析过程通常是从基本统计量的计算和描述开始的，基本的统计分析通常包括单变量频数分布表的编制、基本统计量的计算以及数据的探索性分析等。

通过计算诸如样本均值、样本标准差等重要基本统计量，并辅助以 SPSS 提供的图形功能，能够使分析者把握数据的基本特征和数据的整体分布形态，对进一步的统计推断和数据建模工作起到重要作用。

本章通过例子学习描述性统计分析及其在 SPSS 中的实现，具体内容包括基本描述性统计量的定义及计算、频率分析、描述性分析、探索性分析、交叉表分析和多重响应分析。

3.1 基本描述性统计量简介

描述性统计量是指变量某一特征的统计量，SPSS 提供的基本统计量大致可以分为 3 类：描述集中趋势的统计量、描述离散程度的统计量和描述总体分布形态的统计量。下面分别介绍这 3 类统计量的定义及其计算。

3.1.1 描述集中趋势的统计量

集中趋势是指一组数据向某一中心值靠拢的倾向和程度，计算描述集中趋势的统计量就是要找到能反映数据一般水平的代表值或中心值。统计学中的集中趋势统计量是由样本值确定的值，在频率分布数列中，各观察值有一种向中心集中的趋势，在中心附近的观察值较多，远离中心的较少，这称为集中趋势。它所反映的是一组资料中各种数据所具有的共同趋势，即资料的各种数据所集聚的位置。常用的集中趋势统计量有均值、中位数、众数、总和及百分位数等。

1. 均值

均值（Mean）又称“算术平均值”，指一组数的平均值，其数据定义：

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.1)$$

其中， n 为样本容量； x_i 为样本点的数值。样本均值反映了变量取值的集中趋势或者平均水平，是最常用的基本统计量。

均值适用于数值型数据，不能用于定类数据和定序数据，其缺点是容易受极值或异常值干扰。

2. 中位数

一组样本数据按升序或降序排列后，如果样本容量为奇数，则中位数取中间位置的数值；如果为偶数，则取中间两个数据的平均值。中位数受数据变化影响比均值大，但不易受极值或异常值的干扰。

中位数主要用于定序数据，也可用于数值型数据，但不能用于定类数据。

3. 众数

众数的值是一组数据中出现频数最多的变量值，可能有多个众数，多用于定类数据，也可用于定序数据和数值型数据，不易受极值或异常值干扰。

众数的计算只适用于单位数较多，且存在明显集中趋势的情况，否则众数是没有意义的。

4. 总和

总和表示某变量所有值的和。

5. 百分位数

百分位数类似于随机变量分位点的概念。将样本数据按升序排列后，排在前面 $p\%$ 的数据的右端点的值称为样本的 p 分位数。常用的有四分位数 (Quartile)，指将数据分为四等分，分别位于 25%、50% 和 75% 处的分位数。百分位数 (Percentile Value) 适合于定序数据及度量尺度更高的数据，也可用于数值型数据，不能用于定类数据。

百分位数同中位数一样，不易受极值的影响。

3.1.2 描述离散程度的统计量

离散程度是指一组数据远离其“中心值”的程度。仅仅利用描述集中趋势的统计量，不能反映整个数据集合的分布状况，具有不同分布的数据可能具有相同的平均值、中位数或众数等，因此还需要统计量来反映数据与集中趋势统计量之间的离散程度。统计学中描述离散程度的统计量是样本值远离集中趋势统计量程度的定量化描述，比较重要的描述离散程度的统计量有样本方差、样本标准差、均值标准误差、极差等。

1. 样本方差

样本方差的数学定义：

$$\text{Var} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.2)$$

其中， n 为样本容量； x_i 为样本点的数值。样本方差是刻画样本数据关于均值的平均偏差平方的一个量，是描述样本离散趋势的最常用的统计量。样本方差越大，表明样本偏离样本平均值的可能性越大。

各变量值对均值的方差小于对任意值的方差。

2. 样本标准差

由于样本方差的计算单位是样本值的平方，将样本方差开方后可以得到和样本值相同量纲的统计量，我们将样本方差开方后的统计量称为样本标准差。样本标准差和样本方差一样，也是度量样本离散程度的重要统计量。

3. 均值标准误差

均值标准误差即样本均值的标准差，其数学定义：

$$\text{S.E. Mean} = \frac{\sigma}{\sqrt{n}} \quad (3.3)$$

其中， n 为样本容量； σ 为总体分布的标准差。均值标准误差是描述样本均值和总体值平均偏差程度的统计量。

4. 极差

一种简单的度量数据分散度的方法就是找出极差 (Range)，即最大与最小观察值的差：

极差 = 最大观察值 - 最小观察值

极差很容易计算，而且常常是一个很有用的数。数据的平均值和它的极差可以告诉我们很多被观测变量的信息。如果数据不包含一些极端的值，平均值就会更准确。极差有一个缺点，就是对极端值十分敏感。

现实生活中，可以用极差值来对产品质量进行检验。正常条件下，一组产品质量稳定，极差应该在一定范围内波动，若极差超过给定范围，则说明出现异常。

3.1.3 描述总体分布形态的统计量

集中趋势和离散程度是数据分布的重要特征，但要从整体上全面把握样本数据的分布，仅有集中趋势和离散程度统计量是不够的，还需要掌握数据分布的形态，例如直方图的对称性、偏斜程度以及陡缓程度等。描述数据分布形态的统计量主要有偏度和峰度两种。

1. 偏度

偏度是描述取值分布形态对称性的统计量，偏度的数学定义：

$$\text{Skewness} = \frac{\mu_3}{\text{Var}^{3/2}}$$

(3.4)

其中，

$$\mu_3 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^3$$

Var 为样本方差。偏度系数大于 0，表示其数据分布形态有一条长尾拖在右边，称为右偏或正偏，偏度系数小于 0，表示其数据分布形态有一条长尾拖在左边，称为左偏或负偏。偏度系数的绝对值越大，与正态分布相比越偏斜。来自于正态总体的样本偏度近似为 0。

2. 峰度

峰度是描述变量取值分布形态陡缓的统计量，峰度的数学定义：

$$\text{Kurtosis} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^4 / S^4 - 3$$

(3.5)

其中，S 为样本标准差。当峰度系数等于 0 时，表明数据分布的陡峭程度与正态分布相同；当峰度系数大于 0 时为尖峰分布，表明数据分布的陡峭程度比正态分布大；当峰度系数小于 0 时为扁平峰分布，表明数据分布的陡峭程度比正态分布小。

所以，可以利用偏度和峰度的值是否接近 0 作为检验是否是正态分布的重要依据。

偏态与峰态分布的形状如图 3-1 所示。

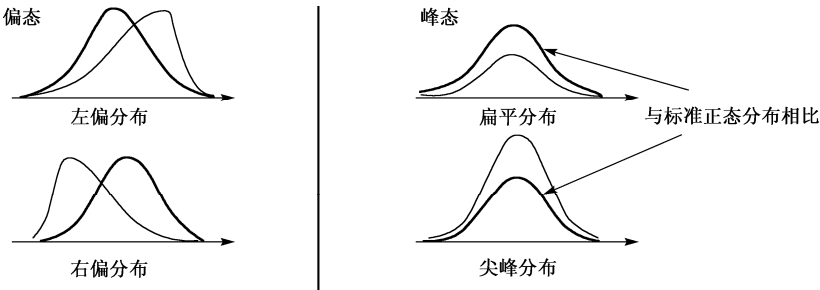


图 3-1 偏态与峰态分布的形状

3.2 频率分析

3.2.1 基本概念及统计原理

频率就是一个变量在各个变量值上取值的个案数，分析时不考虑其实际取值。基本统计分析往往从频率分析开始。通过频率分析能够了解变量取值的状况，对把握数据的分布特征是非常有用的。

例如，调查消费者拥有数码产品的数量，首先分析受访者的总人数、家庭收入情况、受教育程度、性别等，获取样本是否具有总体代表性、抽样是否存在系统偏差等信息。这些可以通过频率分析来实现，经过频率分析可以得到如下结果：

- (1) 频率分布表：该表中包含频率、各频率占总样本数的百分比、有效百分比、累计百分比。
- (2) 统计图：用统计图形展示变量的取值状况，频率分析中提供的统计图形可以是条形图、饼图或直方图。

3.2.2 SPSS 实例分析

【例 3-1】 以下是调查问卷中针对被调查人设置的两个问题。

1. 您的家庭月收入大约是：(请包括所有工资、奖金、津贴等在内，以元为单位) 单选

500 ~ 1000.....1	1000 ~ 1999.....2	2000 ~ 2999.....3	3000 ~ 3999.....4
4000 ~ 4999.....5	5000 ~ 5999.....6	6000 ~ 6999.....7	7000 ~ 7999.....8
8000 ~ 8999.....9	9000 ~ 9999.....10	10000 及以上...11	

2. 您的受教育程度(指您受过的最高或正在接受的教育程度) 单选

没有受过正式教育/小学.....1	初中.....2	高中/中专/技校.....3
大专/大学非本科/高职高专.....4	大学本科.....5	研究生及以上.....6

从问卷中收集到的数据如表 3.1 所示。

试对收集到的数据进行频率分析。(参见数据文件：data3-1.sav。)

对该数据文件中的两个变量进行频率分析的具体步骤如下。

第 1 步 数据组织。

生成的 SPSS 数据文件，建 2 个变量：“收入”、“教育”，保存为数据文件 data3-1.sav。

第 2 步 频率分析设置。

(1) 选择菜单：“分析→描述统计→频率”，打开“频率(F)”对话框，并按图 3-2 所示进行设置。

该对话框主要由以下几部分组成。

- ① 候选变量列表框：图 3-2 左侧的列表框，存放文件中所有的变量。
- ② 变量(V)：存放待分析变量。将要分析的变量从左侧的候选变量列表框移入右侧的“变量(V)”框中。
- ③ 显示频率表：设置是否显示频率表，系统默认选中，表示在分析结果中将显示分析变量的频率分布表。

(2) “统计”选择：确定要输出的统计量。

单击图 3-2 中的“统计(S)…”按钮，出现“频率：统计”对话框，按图 3-3 所示进行设置。

表 3.1 问卷调查结果

收 入	教 育
3	5
1	5
2	5
1	5
1	5
...	...
6	4
10	5
4	5



图 3-2 “频率”对话框

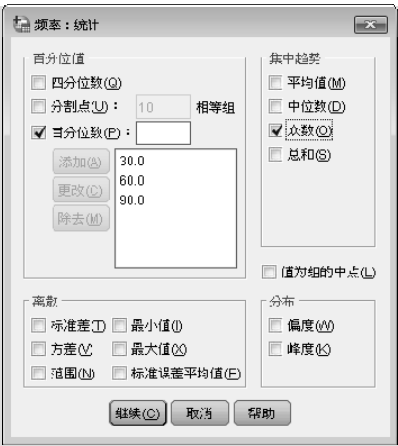


图 3-3 “频率：统计”对话框

该对话框主要由以下几部分组成。

① “百分位值”复选框组：设定在频率分析结果中输出哪些百分位数。

- 四分位数：设置是否显示分析变量的四分位数。
- 分割点：设定将数据平均分为所设定的相等等份，并在结果中显示。
- 百分位数：显示用户自定义的百分位数，在文本框中输入数值的范围为 0~100，输入数值后单击“添加(A)”按钮加入到列表框中，可重复输入。

② “集中趋势”复选框组：设定在频率分析结果中输出哪些集中趋势统计量。“离散”和“分布”复选框组功能类似，其中“离散”复选框组中的“范围”即离散统计量“极差”。

③ “值为组的中点(L)”：分组计算中位数和百分位数。选中该复选框，在计算百分位数和中位数时，如果数据已经分组，就按已经分组的数据计算各组数据的中位数和百分位数。

(3) “图表”选择：确定要输出的统计图形。

单击图 3-2 中的“图表(C)...”按钮，打开“频率：图表”对话框。按图 3-4 所示进行设置。该对话框主要由以下几部分组成。

① “图表类型”单选框组：设置输出的图形类型。

- 无：默认选项，不生成图形。
- 条形图：生成条形图。
- 饼图：生成饼图。
- 直方图：生成直方图。
- 在直方图中显示正态曲线(S)：选中“直方图”后才能选择该项，选中后，分析结果中将输出数据的直方图，并且在直方图上绘制出正态曲线，用于推断数据是否服从正态分布。

② “图表值”单选框组：设定图形的取值。

- 频率：默认选项，选择该项，表示按照频率作图。
- 百分比：按照百分比作图。

(4) “格式”选择：确定要输出的数据格式。

单击图 3-2 中的“格式(F)...”按钮，打开“频率：格式”对话框。按图 3-5 所示进行设置。该对话框主要由以下几部分组成。

① “排序方式”单选框组：定义输出频率表的数据排列次序。

- 按值排序：表示按照数据的升序或降序排列频率分布表。

- 按计数排序：表示按照频数的升序或降序排列频率分布表。
- ② “多个变量”单选框组：设置多个变量的结果是否在同一表格输出。
- 比较变量：将各变量的统计结果在同一统计量表输出。
- 按变量组织输出：将各变量的统计结果在各自的统计量表中输出。
- ③ “排除具有多个类别的表”：表示当频数表的分组大于下面设定的数值时，禁止其在结果中输出，这样可以避免产生巨型表格。

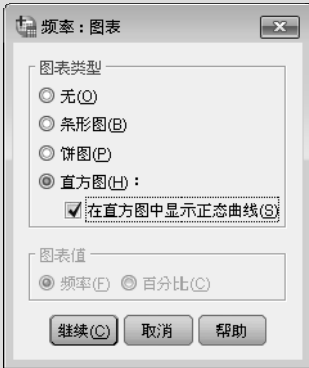


图 3-4 “频率：图表”对话框

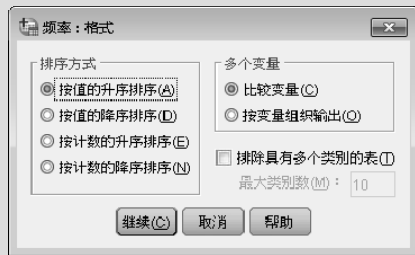


图 3-5 “频率：格式”对话框

第3步 主要结果及分析。

运行结果如表 3.2~表 3.4 及图 3-6、图 3-7 所示，各表和图的具体意义分析如下。

(1) 表 3.2 是两个变量的变量值基本信息：有效的个案数、缺失值个数、众数，百分位数。因为在“格式”设置中选择了“比较变量”，所以“教育”和“收入”两个变量的统计结果显示在同一表中。

(2) 表 3.3 是变量“教育”的频率分布表，即每一个变量值的频率、百分比、有效百分比、累计百分比。从表中可以看出，变量值为 5 的个案数最多，即受访者中受教育程度为“大学本科”的人最多。

表 3.2 统计量表

		收入	教育
个案数	有效	835	835
	缺失	0	0
众数		3	5
百分位数	30	3.00	4.00
	60	4.00	5.00
	90	7.00	5.00

表 3.3 变量“教育”的频率分布表

		频率	百分比	有效百分比	累计百分比
有效	1	8	1.0	1.0	1.0
	2	39	4.7	4.7	5.6
	3	114	13.7	13.7	19.3
	4	165	19.8	19.8	39.0
	5	456	54.6	54.6	93.7
	6	53	6.3	6.3	100.0
	总计	835	100.0	100.0	

(3) 表 3.4 是变量“收入”的频率分布表，即每一个变量值的频率、百分比、有效百分比、累计百分比。从表中可以看出，变量值为 3 的个案数最多，即受访者中家庭收入在“2000~2999”的人最多。

(4) 图 3-6 和图 3-7 是两个变量的直方图，从图上看，受访者受教育程度同正态分布相比右偏，受访者家庭收入的分布左偏，都不具明显的正态分布。

表 3.4 变量“收入”的频率分布表

		频率	百分比	有效百分比	累计百分比
有效	0	2	.2	.2	.2
	1	87	10.4	10.4	10.7
	2	152	18.2	18.2	28.9
	3	156	18.7	18.7	47.5
	4	137	16.4	16.4	64.0
	5	88	10.5	10.5	74.5
	6	85	10.2	10.2	84.7
	7	52	6.2	6.2	90.9
	8	27	3.2	3.2	94.1
	9	9	1.1	1.1	95.2
	0	8	1.0	1.0	96.2
	11	32	3.8	3.8	100.0
总计		835	100.0	100.0	

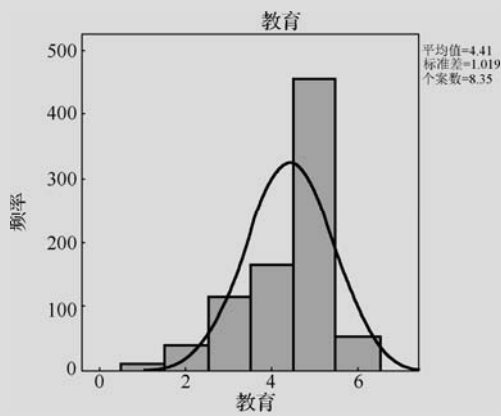


图 3-6 变量“教育”的直方图

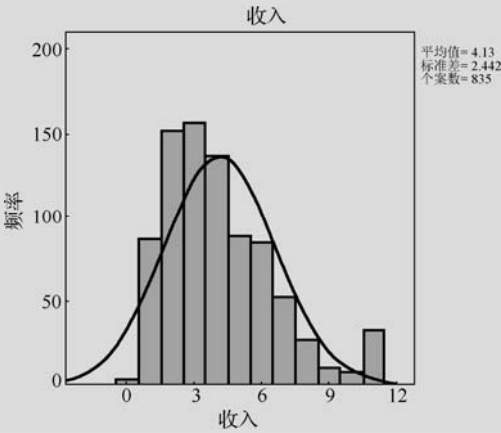


图 3-7 变量“收入”的直方图

3.3 描述性分析

3.3.1 基本概念及统计原理

描述性分析主要用于计算并输出变量的各类描述性统计量的值，通过 3.2 节的学习可知，频率分析同样可以做到，两者都以计算数值型变量的统计量为主。描述性统计分析没有图形功能，也不能生成频率表，但描述性分析可以将原始数据标准化为 Z 分数，在当前数据文件中添加新变量，用于保存相应的 Z 标准分数，其变量名为相应变量名前加字母 Z，以便后续分析时应用。Z 标准分数是一个变量值与该变量平均值之差与标准差的比值，标准化处理后，可以保证数据服从标准正态分布。

Z 变换的公式：

$$Z_i = \frac{x_i - \bar{x}}{S}$$

(3.6)

式中， x_i 是变量的样本值； \bar{x} 是样本均值； S 是样本标准差。

3.3.2 SPSS 实例分析

【例 3-2】 图 3-8 是 5 岁儿童体重、身高、胸围的部分 SPSS 数据，试对儿童身高作描述性统计分析。（参见数据文件：data3-2.sav。）

对该数据文件中的变量进行描述性分析的具体步骤如下。

第 1 步 打开数据文件 data3-2.sav。

第 2 步 描述性分析设置。

（1）选择：“分析→描述统计→描述”，打开“描述”主对话框，确定要进行描述性分析的变量。按图 3-9 所示进行设置。

该对话框主要由以下几部分组成。

① 候选变量框：图 3-9 左侧变量列表框，列出数据文件包含的所有变量。

② 变量：从候选变量框中选择要进行描述性分析的变量移入此框中，可同时选择多个变量，此时，SPSS 就将分别产生多个变量的描述性分析结果。

③ 将标准化值另存为变量：用于确定是否在当前数据文件中生成 Z 标准分数。

（2）“选项”选择：用于确定要输出的统计量；在图 3-9 中单击“选项...”按钮，打开“描述：选项”子对话框，并按图 3-10 所示进行设置。该对话框同频率分析中“统计”对话框中统计量的选择类似，见图 3-3。

体重	身高	胸围
17	110.6	55
15	103.2	50
21	112.5	55
16	106.8	50
18	109.7	56
20	111.1	55
17	105.8	51
17	109.5	53
18	109.2	53

图 3-8 5 岁儿童部分 SPSS 数据



图 3-9 “描述”主对话框



图 3-10 “描述：选项”子对话框

第 3 步 运行结果及分析。

运行结果如表 3.5 所示，该结果包括变量值的个数、极值、均值、标准差、偏度和峰度信息。输出统计量中，方差和标准差越小越好，越小说明该组数据越趋于稳定。

表 3.5 描述性分析输出结果

	个案数	最小值	最大值	平均值	标准差	偏度		峰度	
	统计	统计	统计	统计	统计	统计	标准误差	统计	标准误差
身高	96	99.3	125.0	109.891	5.9633	.350	.246	-.446	.488
有效个案数（成列）	96								

3.4 探索性分析

3.4.1 基本概念及统计原理

探索分析是一种在对资料的性质、分布特点等完全不清楚的情况下，对变量进行更深入研究的描述性统计方法。与前面介绍的两种分析方法相比，探索性分析更加强大大，增加了有关数据文字与图形的描述，可以对变量进行更为深入详尽的统计分析。在进行统计分析前，通常需要寻求和确定适合所研究问题的统计方法，SPSS 提供的探索性分析是解决此类问题的有效办法。

探索性分析提供了很多关于数据的概括分析和图表直观描述的方法，不仅对个案数据有效，而且可以针对分组个案。在输出常用描述性统计量的基础之上，探索性分析增加了有关数据详细分布特征的文字与图形表述，如茎叶图、箱图等，更加详细、完整，还可以提供正态分布检验和方差齐性检验，有助于用户制定进一步分析的方案。

3.4.2 SPSS 实例分析

【例 3-3】 表 3.6 是某班 3 门课程对应成绩的统计数据，试对其进行探索性分析并做是否服从正态分布的检验。（参见数据文件：data3-3.sav。）

表 3.6 某班 3 门课程对应成绩的统计数据

科目	1	1	1	1	1	1	2	2	2
成绩	83	74	73	30	60	95	73	11	16
科目	2	2	2	3	3	3	3	3	3
成绩	75	56	19	85	91	11	55	32	56

- 对该数据文件中的两个变量进行探索性分析和正态分布检验的具体步骤如下。
- 第 1 步 数据组织。**
- 根据表 3.6 生成 SPSS 数据文件，建 2 个变量：“科目”、“成绩”，“科目”的度量标准为“名义”，“成绩”的度量标准为“标度”，数据文件的格式与表 3.6 类似，保存为数据文件 data3-3.sav。
- 第 2 步 探索分析设置。**
- (1) 菜单选择：“分析→描述统计→探索”，打开“探索”对话框，按图 3-11 所示进行设置。该对话框主要由以下几部分组成。
- ① 候选变量框：图 3-11 左侧列表框为候选变量框，列出数据文件中的所有变量。
- ② 因变量列表 (D)：用于存放待分析的变量，可以同时选择多个变量（选择的变量必须是数值型变量）。
- ③ 因子列表：用于选择分组变量，根据该变量的不同取值，分组分析“因变量列表”中的变量，可以没有因子变量，也可有多个因子变量。
- ④ 个案标注依据：用于选择标签变量，只能选一个，该文本框中的变量作为标识符，在输出诸如异常值时，用该变量进行标识，如果该项缺选，则系统自动寻找“id”变量作为变量标签。
- ⑤ 输出：用于设定在分析结果中输出的内容。
- 两者：同时输出统计量和图；
 - 统计：只输出统计量，不输出图；
 - 图：只输出图，不输出统计量。

(2) “统计”选择：确定探索性分析结果中将要输出的统计量。

单击图 3-11 中的“统计...”按钮，打开“探索：统计”对话框，按图 3-12 所示进行设置。该对话框主要由以下几部分组成。

① “描述”复选框：系统默认选项，用于输出基本的描述性统计量，包括均值、中位数、5%的调整均值、标准误差、极差、最大值、最小值、四分位数、峰度和偏度及其标准误差等。

选择该项时，需要在下方“平均值的置信区间”文本框中输入 1%~99%间的任意值，系统根据该值计算出置信区间的上下限，系统默认为 95%。



图 3-11 “探索”对话框



图 3-12 “探索：统计”对话框

② “M-估计量”复选框：用于输出 4 种稳健极大似然估计量，对于长尾对称分布或数据有极端异常值时，利用稳健估计量估计总体均值比用样本均值或中位数有更好的稳定性；根据样本值的权重不同，可以得到不同的估计量，主要有 4 种，包括 Huber（稳健估计量）、Hample（非降稳健估计量）、Andrew（波估计量）、Turkey（复权估计量）。

③ “离群值”复选框：用于输出数据的离群点，将输出 5 个最大值和 5 个最小值，输出在窗口中加以显示。

④ “百分位数”复选框：用于输出百分数，包括 5%、10%、25%、30%、75%、90%和 95%的百分位数。

(3) “图...”选择：用于确定探索性分析输出的统计图形。

单击图 3-11 中的“图(T)...”按钮，打开“探索：图”对话框，按图 3-13 所示进行设置。

该对话框主要由以下几部分组成。

① “箱图”选项组：设置显示箱图。

➤ “因子级别并置”：为每个因变量生成一个箱图，用于比较同一因变量在分组变量值的不同水平上值的分布情况，同一因变量的不同分组显示在同一个箱图中。

➤ “因变量并置”：所有因变量生成一个箱图，用于比较同一分组水平下不同因变量的值的分布，同一分组的不同因变量显示在同一个箱图中。

➤ “无”：不显示箱图。

② “描述图”选项组：用于设置图形输出时图形的种类，选中表示输出结果中将显示对应图形，可不选择。

③ “含检验的正态图”：选中此项，表示将进行正态分布检验，并生成标准 Q-Q 图和趋降



图 3-13 “探索：图”对话框

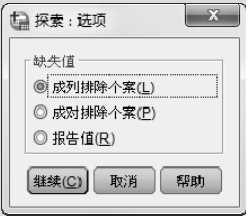
标准 Q-Q 图,同时输出 K-S 统计量中的显著性水平检验,如果观测数目不超过 20,将用 S-W 统计量代替 K-S 统计量。

④ “含莱文检验的分布-水平图”选项组:用于对数据转换所得散布水平图的设置。对于所有的散布水平图,显示数据转换后回归曲线的斜率和方差齐性的 Levene 稳健检验,有如下 4 项设置。

- 无:系统默认选中,表示不进行方差齐性检验。
- 幂估算:对每组数据产生一个中位数范围的自然对数与四分位范围的自然对数的散点图,同时在每组中数据方差相等的条件下对数据进行幂变换的估计。
- 转换后:对原始数据进行转换,在“幂”下拉菜单中选择幂变换使用的幂值。选择转换函数后,可以产生转换后的数据散布图。
- 未转换:原始数据不进行转换。

(4) “选项”选择:用于确定分析过程中对缺失值的处理方式。

单击图 3-11 中的“选项(O)…”按钮,打开“探索:选项”对话框,按图 3-14 所示进行设置。



该对话框主要由以下几部分组成。

- ① 成列排除个案:该选项是系统的默认选项,表示去除所有含缺失值的个案后再进行分析;
- ② 成对排除个案:去除当前分析变量中有缺失值的个案及与缺失值有成对关系的个案;
- ③ 报告值:将分组变量的缺失值单独分为一组,并在频数表中输出。

图 3-14 “探索:选项”对话框

本例中直接使用默认设置。

第 3 步 运行结果及分析。

完成以上操作步骤后,单击图 3-11 中的“确定”按钮,运行结果如表 3.7~表 3.10 及图 3-15~图 3-17 所示,具体意义分析如下。

① 表 3.7 是探索性分析的数据摘要,很多 SPSS 统计分析过程会自动给出一个这样的数据摘要,表中给出参与分析的变量或变量分组的个案数、缺失信息等。在本例中,每个变量分组有 6 个个案参与分析,无缺失值。

表 3.7 个案处理摘要

	科目	个案					
		有效		缺失		总计	
		个案数	百分比	个案数	百分比	个案数	百分比
成绩	语文	6	100.0%	0	0.0%	6	100.0%
	数学	6	100.0%	0	0.0%	6	100.0%
	英语	6	100.0%	0	0.0%	6	100.0%

② 表 3.8 中输出的是描述性统计量。在本例中,输出了“成绩”按“科目”分组的各组描述性统计量,除了 3.1 节介绍的统计量之外,表 3.8 还多生成了几个特殊的统计量,分别是均值的 95%置信区间、5%修整均值、四分位距(即 3/4 分位点与 1/4 分位点之差)。

③ 表 3.9 给出了数据的 M 估计值。在 SPSS 中,根据权重系数的不同,共提供了 4 种估计方法,表 3.9 下方的注释分别给出了 4 种方法的加权常量。通常,对于有异常或极端值的数据,M 均值估计法有很好的稳定性,用 M 估计值替代均值或中位数,结果更准确。因此,如果探索

性分析中描述性统计量中的均值和 M 均值有较大的差距，那么用户就应当注意数据中是否有异常值了。

表 3.8 描述性统计量表

描述				
科目			统计	标准误差
成绩	语文	平均值	69.17	9.156
		平均值的 95% 置信区间	下限	45.63
			上限	92.70
		5% 剪除后平均值	69.91	
		中位数	73.50	
		方差	502.967	
		标准差	22.427	
		最小值	30	
		最大值	95	
		全距	65	
		四分位距	34	
		偏度	-1.085	.845
		峰度	1.617	1.741
数学		平均值	41.67	12.126
		平均值的 95% 置信区间	下限	10.50
			上限	72.84
		5% 剪除后平均值	41.52	
		中位数	37.50	
		方差	882.267	
		标准差	29.703	
		最小值	11	
		最大值	75	
		全距	64	
		四分位距	59	
		偏度	.153	.845
		峰度	-2.812	1.741
英语		平均值	55.00	12.466
		平均值的 95% 置信区间	下限	22.96
			上限	87.04
		5% 剪除后平均值	55.44	
		中位数	55.50	
		方差	932.400	
		标准差	30.535	
		最小值	11	
		最大值	91	
		全距	80	
		四分位距	60	
		偏度	-.250	.845
		峰度	-1.002	1.741

表 3.9 M 均值估计表

科目		休伯 M 估计量 a	图基双权 b	汉佩尔 M 估计量 c	安德鲁波 d
成绩	语文	72.54	75.78	72.88	76.05
	数学	41.42	41.13	41.67	41.13
	英语	56.68	55.60	55.00	55.60

- a. 加权常量为 1.339。
- b. 加权常量为 4.685。
- c. 加权常量为 1.700、3.400 和 8.500。
- d. 加权常量为 1.340*pi。

④ 表 3.10 是探索性分析的正态性检验结果表。分别利用柯尔莫戈洛夫-斯米诺夫 (Kolmogorov-Smirnov) 检验和夏皮洛-威尔克 (Shapiro-Wilk) 检验两种方法来确定变量是否服从正态分布。一般来说, 显著性 P 值大于 0.05 表示变量服从正态分布的显著性强。正常情况下, 两种方法的检验结论应该一致; 某些时候, 当以上两种检验方法结论矛盾时, 大样本以 K-S (Kolmogorov-Smirnov) 检验为准, 小样本以 S-W (Shapiro-Wilk) 检验为准 (样本数 < 30 为小样本)。

本例中, 三个分组的两种方法显著性 P 值均大于 0.05, 因此三个分组均服从正态分布。

表 3.10 正态性检验表

科目		柯尔莫戈洛夫-斯米诺夫 a			夏皮洛-威尔克		
		统计	自由度	显著性	统计	自由度	显著性
成绩	语文	.235	6	.200*	.929	6	.573
	数学	.277	6	.165	.827	6	.102
	英语	.170	6	.200*	.946	6	.706

*. 这是真显著性的下限。
a. 里利氏显著性修正。

⑤ 图 3-15 所示为成绩按科目分组后各分组的箱图, 由于在分析过程中的“图形”对话框中选择的是“因子级别并置”, 所以成绩按科目分成的三个分组的箱图绘制在同一张图上。每一个箱体上方那条线的取值代表最大值, 下方那条线的取值代表最小值。箱体自身的三条线从上到下分别代表 3/4 分位点、中位点、1/4 分位点的取值。从图中还可以看出, 成绩的三个分组中均没有离群值, 若有离群值则会在图中用 “.” 标注出来。

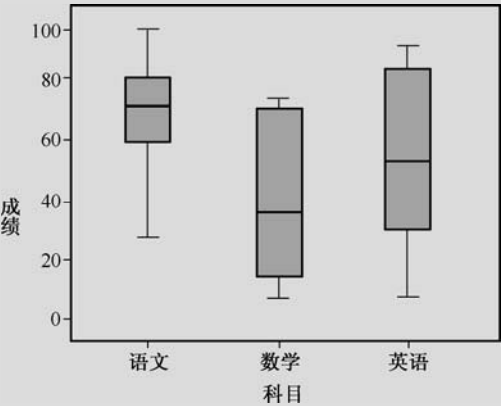


图 3-15 箱图

⑥ 探索性分析的标准 Q-Q 图和趋降标准 Q-Q 图。

在前面的章节中介绍了直方图，直方图可以大致判断数据满足的分布类型，但是这种判断完全依赖于统计工作者的实际经验，难免有偏差。标准 Q-Q 图可以检验数据是否服从某种分布，在标准 Q-Q 图中，检验数据是否较好地服从给定分布的标准有两个：①看标准 Q-Q 图上的数据点与直线的重合度；②趋降标准 Q-Q 图上的点是否关于直线 $Y=0$ 在较小的范围内上下波动。

探索分析中生成的标准 Q-Q 图以及趋降标准 Q-Q 图用于检验数据是否服从正态分布，在本例中，变量“成绩”按科目“语文”、“英语”和“数学”分组，分别生成标准 Q-Q 图、趋降标准 Q-Q 图，图 3-16 和图 3-17 为成绩按科目“语文”分组的标准 Q-Q 图、趋降标准 Q-Q 图，从图中可以看出，两个变量的数据都很好服从正态分布，这也和表 3.10 的检验结果相吻合。

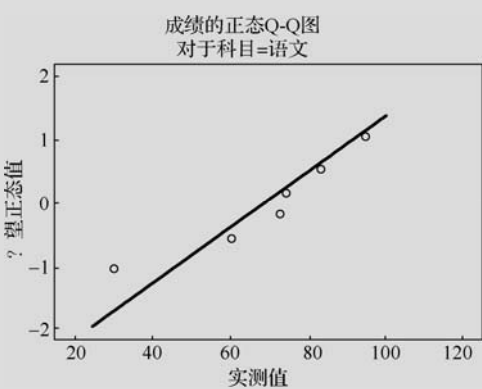


图 3-16 科目=语文的成绩标准 Q-Q 图

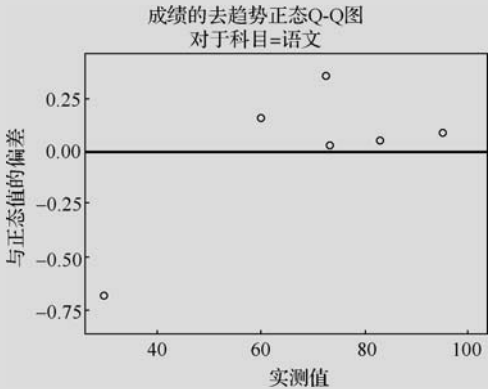


图 3-17 科目=语文的成绩趋降标准 Q-Q 图

3.5 交叉表分析

3.5.1 基本概念及统计原理

1. 交叉表分析的概念

前面学习的频率分析、描述性分析及探索分析，都是针对单变量自身的数据分布情况进行分析的。在实际分析中，常需要分析多个变量之间，一个变量是否对其他变量的取值存在影响，分析变量之间是否存在关系，这种分析就称为交叉表分析。在分析变量之间的关系时，通常分析变量之间的相关程度。对于数值型变量，分析其相关性通常是计算相关系数或进行回归分析，这在后面的章节中有较为详细的介绍。而对于定类型变量，则通常采用交叉表进行分析。

交叉表是两个或多个变量交叉分组后形成的频数分布表，主要用于研究定类型变量之间有无相关性，给出了变量在不同取值下的数据分布。交叉表分析根据样本数据，产生二维或多维交叉表，并在产生交叉表的基础上，对两两变量间是否存在一定的相关性进行分析。

交叉表分析法的应用极为广泛，它可以分析研究总体中个体的属性之间是否相关，称为独立性检验。

2. 交叉表分析的相关关系的主要检验方法

在分析中，难以在交叉表中直接发现行、列变量之间的关系及关系强度，需要借助非参数检验方法和度量变量间相关程度的统计量进行分析，通常采用 χ^2 检验和相关性检验。

在交叉表分析中，SPSS 提供的相关关系的检验方法主要有以下几种。

(1) 卡方统计检验：常用于检验行列变量之间是否相关。计算公式为

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

(3.7)

式中， f_0 为实际观察频数； f_e 为期望频数。

卡方统计量服从（行数1）×（列数1）个自由度的卡方统计。SPSS 在计算卡方统计量后，会给出相应的相伴概率，通过比较相伴概率及显著性水平，来判断行列变量之间是否相关。

(2) 列联系数：用于名义变量之间的相关系数计算。计算公式为

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

(3.8)

式中， C 为列联系数； N 为样本数。

由式 (3.8) 可见，列联系数 C 由卡方统计量修改而来，其取值范围为 0~1。 C 越接近 1，表明卡方值足够大而使得样本数起的作用极小，应拒绝原假设，认为行列变量有较强的相关关系； C 越接近 0，表明卡方值足够小而使得样本数起的作用极大，因此不应拒绝原假设。

(3) V 系数 (Phi and Cramer’s V)：常用于名义变量之间的相关系数计算。计算公式为

$$V = \sqrt{\frac{\chi^2}{N(K-1)}}$$

(3.9)

式中， N 为样本数； K 为行数和列数较小的实际数。

由式 (3.9) 可见， V 系数也由卡方统计量修正而来，在考虑了样本数影响的同时，还考虑了交叉表的单元格数，其取值范围为 0~1。

3.5.2 SPSS 实例分析

【例 3-4】 在设置学生评价实验教学的调查表中，“实验准备”是其中的一项指标，为分析“实验准备”情况与评价结果的关系，建立的 SPSS 数据文件中的部分数据如图 3-18 所示，变量值标签如表 3.11 所示。（参见数据文件：data3-4.sav。）

实验准备	评价结果
3	3
2	1
2	3
1	1
3	2
2	1
3	2
2	1
1	1
2	1

表 3.11 变量值标签

变量	实验准备	评价结果
值与值标签	1	差
	2	一般
	3	准备充分
		1 差
		2 一般
		3 优

图 3-18 数据文件

对该数据文件中的两个变量进行交叉表分析的具体步骤如下。

第 1 步 数据组织。

在数据文件中建立两个变量：“实验准备”、“评价结果”，两个变量均为数值型或字符型的定类变量，其度量标准为“名义”，并根据表 3.11 定义各变量的变量值标签，保存为 SPSS 数据文件 data3-4.sav。

第2步 交叉表分析设置。

(1) 选择菜单：“分析→描述统计→交叉表”，打开“交叉表”对话框，按图 3-19 所示进行设置。



图 3-19 “交叉表”对话框

该对话框主要由以下几部分组成。

- ① 候选变量框：对话框左侧的列表框，列出数据文件中所有的变量；
 - ② “行”、“列”文本框：分别用于选择交叉表的行、列变量，行、列变量必须是数值型或字符型等定类变量；
 - ③ 层 1/1：用于选择分层变量，如果除行、列变量外，还有其他变量参加分析，则将其加入到“层 1/1”列表框中，用于生成多维交叉表，用“上一个”、“下一个”按钮控制分层的层数；
 - ④ “显示簇状条形图”复选框：用于确定是否在输出文件中显示簇状条形图；
 - ⑤ “排除表”复选框：用于确定是否在输出文件中显示分析结果的“交叉制表”；
- (2) “精确”选择：该对话框提供了三种不同条件的检验方式来检验行、列变量的相关性。单击图 3-19 中的“精确(X)...”按钮，打开“精确检验”对话框，按图 3-20 所示进行设置。

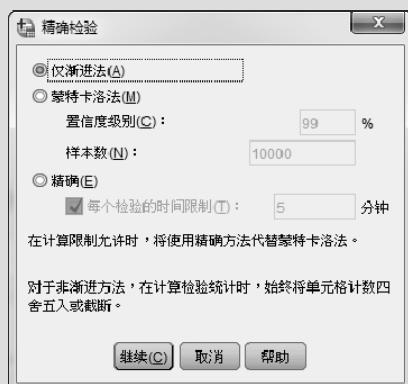


图 3-20 “精确检验”对话框

该对话框主要由以下几部分组成。

- ① “仅渐进法”：为默认选项，适用于具有渐进分布的大样本数据，不适用于小样本和非渐进分布方式的检验；
 - ② “蒙特卡洛法”：用于指定个案数量的检验，该选项允许非渐进分布方式的检验，需设置“置信水平”和“样本数”，置信区间一般为 90、95、99；
 - ③ “精确”：为精确计算，需要设置每次精确检验所花费的最大时间限度，若超过 30 分钟，则使用“蒙特卡洛法”。
- 一般情况下，此对话框的选项使用系统默认值，本例中使用系统默认值。
- (3) “统计”选择：用于确定检验方法及要输出的统计量。

单击图 3-19 中的“统计 (S) ...”按钮，打开“交叉表：统计”对话框，并按图 3-21 所示进行设置。

该对话框主要由以下几部分组成。



图 3-21 “交叉表：统计”对话框

- ① “卡方”复选框：用于对行、列变量的独立性进行卡方检验，包括皮尔逊卡方、似然比、线性和线性组合 3 种检验结果。这几种检验的作用是不同的，皮尔逊卡方常用在二维表中对行变量和列变量进行独立性假设检验，似然比用于对数据线性模型的检验。
- ② “相关性”复选框：用于选择是否计算相关系数，检验两个变量的线性相关程度，包括 Pearson 相关系数和 Spearman 相关系数两种检验结果。
- ③ “名义”选项组：用于定义定类变量的相关性指标，共有 4 个指标。
 - “列联系数”：表征变量之间相关性的强弱，取值在 0~1 之间，取值为 0 表示行和列变量之间不相关；其值越靠近 1，表示两变量间的相关性越强。
 - “Phi 和克莱姆 V”：用来刻画变量之间的 Phi 相关性。在不同的卡方检验中，取值范围不同，但是指标的绝对值越大，变量间的相关性越强。
 - “Lambda”：反映自变量对因变量的预测效果。取值为 1 表示自变量可以很好地预测因变量，为 0 表示自变量和因变量之间没有可预测的关系。
 - “不确定性系数”：以熵为标准反映一个变量对另一个变量的确定程度。取值为 1 表示可由一个变量的信息完全确定另一个变量的信息，为 0 表示两个变量之间的信息没有关系。
- ④ “有序”选项组：用于定义定序变量的相关性系数，包括以下 4 个指标。
 - Gamma 系数：反映两个有序变量间的对称相关性，取值在-1~1 之间，取 1 或-1 代表两个变量完全一致或不一致，取值为 0 表示两个变量完全不相关。
 - 萨默斯系数：取值在-1~1 之间，结果解释与 Gamma 系数一样。
 - 肯德尔 tau-b 系数：取值在-1~1 之间，结果解释与 Gamma 系数一样。
 - 肯德尔 tau-c 系数：取值在-1~1 之间，结果解释与 Gamma 系数一样。
- ⑤ “按区间标定”选项组：适用于一个名义变量和一个等距变量的相关性检验。
 - “Eta”：反映行列变量的关联程度，取值在 0~1 之间，越接近 1 表示变量的关联程度越高，越接近 0 表示关联程度越低。
- ⑥ “Kappa”复选框：用于设定 Kappa 系数，检验两个评估人对同一对象进行评估时是否具有相同的态度，系数为 1 表示二者态度完全相同，为 0 表示两种评估没有共同点。Kappa 系数只适用于正方表，即两个变量有相同数量的分类。
- ⑦ “风险”复选框：用于设定相对风险比率系数，检验某件事发生和某因子之间的关联性，

此统计量的置信区间包含 1，表示因素与事件无关联。

⑧ “麦克尼马尔”复选框：主要用于检验配对的资料，相对于配对卡方检验；在“验前-验后”的因素设计中，该检验对探测由于实验干扰而产生反应的变化十分有效。

⑨ “柯克兰和曼特尔-亨塞尔统计”复选框：用于进行一个二值因素变量和二值响应变量的独立性检验和齐次性检验，在下面的“检验一般比值比等于”文本框中只能输入正数，系统默认值为 1。

本例中选择“卡方”复选框，对两变量的独立性进行卡方检验。

(4) “单元格”选择：用于指定要输出的统计量。

单击图 3-19 中的“单元格(E)…”按钮，打开“交叉表：单元格显示”对话框，按图 3-22 所示进行设置。

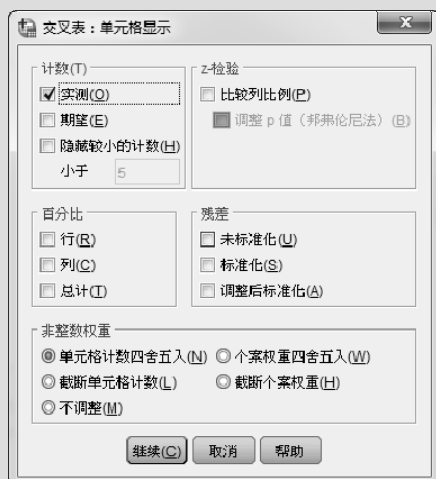


图 3-22 “交叉表：单元格显示”对话框

该对话框主要由以下几部分组成。

① “计数”选项组：用于选择交叉表单元格中的频数输出格式，包含三个选项。

- 实测：显示实际观察值频数，此项是系统默认的选项。
- 期望：输出为理论值。
- 隐藏较小的计数：当期望值小于设定的值时，在输出结果中不显示出来。

② “百分比”选项组：用于选择交叉表单元格中的百分比显示格式，包括三个选项。

- 行(R)：以行为单位，统计行变量的百分比。
- 列(C)：以列为单位，统计列变量的百分比。
- 总计：行、列变量的百分比都进行输出。

③ “残差”选项组：用于选择交叉表单元格中的残差显示格式，包括三个选项。

- 未标准化：单元格中的观测值与预测值之差。
- 标准化： $(\text{观测值} - \text{预测值}) / \text{观测值}$ 。
- 调整后标准化： $(\text{观测值} - \text{预测值}) / \text{标准差}$ 。

④ “非整数权重”选项组：用于当频数因为加权而变成小数时，有以下 5 种调整频数的方法。

- 单元格计数四舍五入：对频数进行四舍五入取整。
- 个案权重四舍五入：对加权样本在使用前进行四舍五入取整。
- 截断单元格计数：对频数进行舍位取整。

- 截断个案权重：对加权样本在使用前舍位取整。
- 不调整：不对计数数据进行调整。
- 本例中使用默认的“观察值”及默认的“四舍五入单元格计数”选项。

第 3 步 主要结果及分析：在执行了以上操作步骤后，运行结果如表 3.12～表 3.14 和图 3-23 所示，分别解释如下。

（1）表 3.12 为案例处理摘要，给出了数据基本信息，表中给出了参与分析的个案数、缺失信息等。在本例中，每个变量有 50 个个案参与分析，无缺失值。

表 3.12 个案处理摘要

	个案					
	有效		缺失		总计	
	N	百分比	N	百分比	N	百分比
实验准备 * 评价结果	50	100.0%	0	0.0%	50	100.0%

（2）表 3.13 给出了数据的 3×3 交叉表，与原始数据在形式上基本一致。

表 3.13 交叉表（实验准备 * 评价结果 交叉制表）

	评价结果			总计
	差	一般	优	
差	12	3	0	15
实验准备 一般	9	8	1	18
准备充分	0	13	4	17
总计	21	24	5	50

（3）表 3.14 是行、列变量通过卡方检验给出的独立性检验结果，共使用了三种检验方法。表下方的注释主要用于决定选择何种卡方检验方法。从表 3.14 可知，各种检验方法显著水平都远小于 0.05，“实验准备与评价结果是独立的”不具有显著性，即认为实验准备这一评价指标与评价结果是相关的。

表 3.14 卡方检验结果

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	22.907a	4	.000
似然比	29.897	4	.000
线性关联	20.357	1	.000
有效个案数	50		

a. 3 个单元格（33.3%）的期望计数小于 5。最小期望计数为 1.50。

在本例中各类卡方检验的结果是一致的，所以避免了选择何种检验方法这一问题。在实际问题中，对于检验方法的选择是不能回避的。通常，交叉表中不应有期望频数小于 1 的单元格，或不应有大量的期望频数小于 5 的单元格，如果交叉表中 20%以上单元格的期望频数小于 5，则不宜使用皮尔逊卡方检验。从皮尔逊卡方统计量的数学定义中可知，如果期望频数偏小的单元格大量存在，皮尔逊卡方统计量无疑会存在偏大的趋势，会易于拒绝原假设。在这种情况下，可以采用似然比检验等方法进行修正。

(4) 图 3-23 的各组状况条形图相当于表 3.13 的直观表示, 用图形表示可直观地得出各种情况的比较。

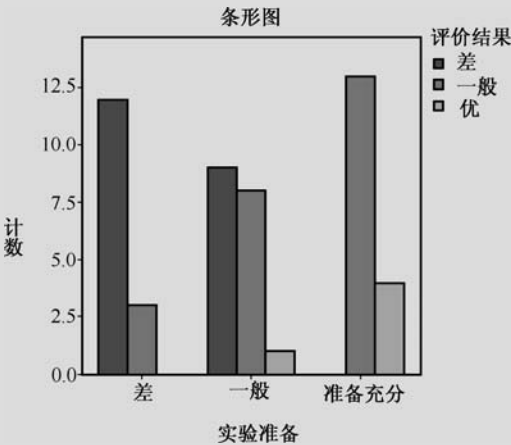


图 3-23 各组状况的分组条形图

3.6 多重响应分析

3.6.1 基本概念及统计原理

1. 基本概念

多重响应分析（即多选项分析）是对多选项问题的分析方法。多选项问题要求问题的答案都是序号变量或名义变量，并且允许选择的答案可以有多个。多选项问题在问卷调查中普遍存在，要求被调查者从问卷给出的若干个可选答案中选择一个以上答案。

例如，调查消费者拥有数码产品的种类，有如下选项：

①数码相机②数码摄像机③MP3④DVD 机

显然，该问题的可选答案有一个以上。通常，在 SPSS 中处理此类问题包括以下两大步骤：

- (1) 将多选项问题分解；
- (2) 利用频率分析或列联表分组下的频率分析方法进行分析。

2. 多重响应问题的分解方法

SPSS 变量中每一个变量值只能保存一个值，无法处理多选项问题中多个答案的问题，对于此类问题，SPSS 的处理方法是 将一个多选项问题分解成若干个问题，对应设置若干个 SPSS 可识别的变量，分别存放描述这些问题的几个可能被选择的答案。这样，对一个多选项问题的分析就可以转化成对多个问题的分析，也就是对多个 SPSS 变量的分析。可见，多选项问题的分解在进行数据分析中是非常关键的。

多选项问题的分解通常有以下两种方法。

(1) 多选项二分法（Multiple Dichotomies Method）

多选项二分法将多选项问题中的每个答案视为一个 SPSS 变量，每个变量只取 0 或 1，分别表示选择该答案或没有选择该答案。

表 3.15 多选项问题的二分分解方法

SPSS 变量名	变量名标签	变量值
V1	数码相机	0/1
V2	数码摄像机	0/1
V3	MP3	0/1
V4	DVD 机	0/1

例如，对于 3.6.1 节中提到的调查消费者拥有数码产品种类的例子，根据提供的答案，对应设置 4 个 SPSS 变量，其取值为 1 或 0，1 表示拥有该产品，0 表示没有该产品，具体如表 3.15 所示。

如果消费者拥有数码相机和 MP3，则变量 V1 和 V3 取值为 1，其余变量取值为 0，经过这样的分解后，就为以后的统计分析做好了准备。

(2) 多选项分类法 (Multiple Category Method)

多选项分类法分解的基本思想是估计多选项问题最多可能出现的答案个数，然后为每个答案定义一个 SPSS 变量值，变量取值为多选项问题中的可选答案。

例如，某部门推选优秀，有 5 个候选人，分别为

- ① 郑一 ② 王二 ③ 张三 ④ 李四 ⑤ 赵五

从这 5 个候选人中推选 3 个，可设置 3 个 SPSS 变量，分别表示优秀人选一、优秀人选二、优秀人选三，变量取值 1~5，依次对应 5 个候选人，具体如表 3.16 所示。

如果某人推选了王二、李四、赵五，则 3 个变量的取值分别对应 2、4、5。

表 3.16 多选项问题的分类分解方法

SPSS 变量名	变量名标签	变量值
V1	优秀人选一	1/2/3/4/5
V2	优秀人选二	1/2/3/4/5
V3	优秀人选三	1/2/3/4/5

3.6.2 多重响应分析 SPSS 实例分析

【例 3-5】对 50 个消费者进行调查，调查其拥有数码产品的种类，有如下选项：

- ①数码相机 ②数码摄像机 ③MP3 ④DVD 机

可多选，试按性别统计拥有各种数码产品的数量。(参见数据文件：data3-5.sav.)

这是一个多选项问题，按照处理多选项问题的方法，分为以下两大步骤。

第 1 步 分解多选项问题，定义多选项变量集。

(1) 分解多选项

按照二分法分解多选项问题，表 3.17 为此多选项问题的二分法记录表，其中“性别”选项 1 为男性，2 为女性，其他选项中的 1 表示拥有该产品，0 表示没有。根据该表建立 SPSS 数据文件，所建立变量的标签与表 3.17 的表头相同，保存为文件 data3-5.sav。

表 3.17 消费者数码产品调查二分表 (前 10 组数据)

编号	性别	数码相机	数码摄像机	MP3	DVD 机
1	1	1	1	1	1
2	1	0	0	1	1
3	2	0	0	0	1
4	1	1	1	0	0
5	1	0	1	1	0
6	1	0	0	0	1
7	2	1	1	1	1
8	2	0	0	0	0
9	2	0	0	1	1
10	1	0	1	0	0

(2) 定义多选项变量集

① 打开数据文件 data3-5.sav。

② 选择菜单“分析→多重响应→定义变量集”，打开“定义多重响应集”对话框，按图 3-24 所示进行设置。



图 3-24 “定义多重响应集”对话框

该对话框主要包括以下几部分。

- 候选变量框：对话框左侧的列表框，列出数据文件中所有的变量。
- 集合中的变量：用于存放要加入集合中的变量；从最左边的候选变量框中选择要加入集合中的变量，添加到“集合中的变量”列表框中。
- 变量编码方式：用来选择变量分解方法。“二分法”选项表示只有两种分类，“类别”选项表示有多种分类。默认的变量分解方法为二分法。

计数值表示输入要进行分析的组号。

- 名称：为多选项变量集命名。
- 标签：对多选项变量集变量名进行说明，改选项可选。

③ 做好以上准备后，“添加（A）”按钮被激活，单击该按钮，将定义好的数据集添加到“多重响应集”列表中，系统会自动在多项集的名称前加字符\$。本例在该列表中将出现名为“\$dp”的多变量数据集名称。

定义好变量集以后，菜单“分析→多重响应”中的“频率（F）...”和“交叉表（C）...”两项才能被激活，如图 3-25 所示。此两项菜单分别表示多重响应下的频率分析及多重响应下的交叉分组频率分析。

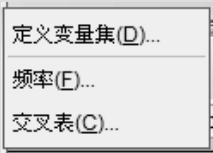


图 3-25 “多重响应分析”子菜单

第 2 步 进行多重响应交叉分组下的交叉表分析。

选择菜单：“分析→多重响应→交叉表”，弹出“多响应交叉表”对话框，按图 3-26 所示进行设置。

该对话框主要由以下几部分组成。

- ① 候选列表框：存放文件中的所有字段。

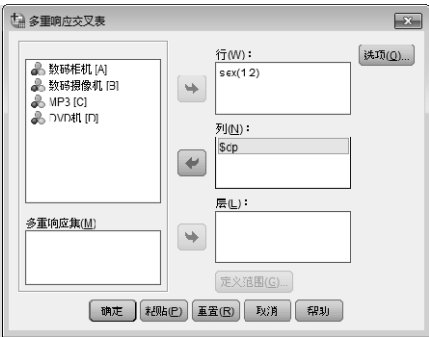


图 3-26 “多重响应交叉表”对话框

- ② “多重响应集”列表框：存放的是步骤 1 中产生的多选项集；选中一个多选项集（本例中为“\$dp”），单击向右箭头按钮，将其添加到“列（N）”文本框中。
- ③ 行（W）：存放交叉表的行变量，添加到“行（W）”列表框中作为交叉表的行，再单击“定义范围（G）...”按钮定义范围，在弹出的对话框中设定最小值为 1，最大值为 2。
- ④ 列（N）：存放作为交叉表的列变量。

第 3 步 主要结果及分析。

经过前两步的分析，运行结果如表 3.18 和表 3.19 所示，分别解释如下。

（1）表 3.18 是多重响应分析的个案摘要表，表中给出了参与分析的个案数和缺失值的信息。

表 3.18 多重响应分析个案摘要

	个案					
	有效		缺失		总计	
	个案数	百分比	个案数	百分比	个案数	百分比
sex*\$dp	45	90.0%	5	10.0%	50	100.0%

（2）表 3.19 是多重响应交叉表分析的结果表。从表中可以看出，男性拥有数码产品的数量高于女性，各种数码产品中，拥有 MP3 的人数最多。

表 3.19 多重响应交叉表分析结果表

		数码产品 a				总计
		数码相机	数码摄像机	MP3	DVD 机	
sex	男 计数	15	16	22	13	30
	女 计数	9	7	13	8	15
总计	计数	24	23	35	21	45

百分比和总计基于响应者。
a. 使用了值 1 对二分组进行制表。

3.7 典型案例

3.7.1 城市平均气温基本特征分析

为了研究城市平均气温的基本特征，现记录了天津及济南两个城市 2007 年 12 个月的平均气温，数据中有 24 个观测样本，分别代表两个城市的 12 个月份；有 3 个属性变量：Month（月份）、

Tj（天津）、Jn（济南），具体数据如表 3.20 所示。（数据来源：杨维忠 等，《SPSS 统计分析 with 行业应用案例详解》，清华大学出版社；参见数据文件：data3-6.sav。）

案例分析：要了解一批数据的基本特征，可以从这批数据的离散趋势、集中趋势及分布形态来入手，现研究两个城市平均气温的基本特征，可以分别求出每个城市的平均气温、众数、最大值、最小值、方差、标准差、峰度、偏度等，通过对天津及济南对应的数据分别进行频率分析、描述性分析或探索分析可以求出。

表 3.20 天津、济南两个城市 2007 年各月份平均气温（单位：℃）

Month（月份）	Tj（天津）	Jn（济南）
1	-2.8	0.0
2	3.3	7.0
3	5.9	8.8
4	14.7	16.0
5	22.0	23.3
6	25.8	26.2
7	27.2	26.6
8	26.4	25.4
9	22.1	21.8
10	13.2	14.7
11	5.6	8.3
12	0	2.3

3.7.2 商场电视品牌满意度调查

为了掌握各品牌电视机的客户认知状况，某商场对 6 种品牌的电视机进行消费者满意度调查，随机调查了 20 位消费者，让他们选出 3 种最满意的电视机品牌，具体数据如表 3.21 所示，答案中的 1~6 分别代表康佳、长虹、西湖、TCL、东芝、创维 6 种品牌。（数据来源：余建英等，《数据统计分析与 SPSS 应用》，人民邮电出版社；参见数据文件：data3-7.sav。）

案例分析：本案例是对最满意电视机品牌问题的调查，数据文件 data3-7.sav 中问题的答案都是名义变量，并且选择的答案有 3 个，是一个多选项问题，所以采取多重响应分析；本题的数据组织形式采用 3.6.1 节中的多选项分类法。

首先定义变量集，将答案 1、答案 2、答案 3 定义为一个变量集，然后进行多重响应分析中的“频率”分析，分析出每种品牌被选择的频率，以及在所有选择中所占的百分比；再采用多重响应分析中的“交叉表”分析，分析出不同性别的消费者对不同电视机品牌选择的情况。

表 3.21 20 名消费者电视机品牌满意度调查情况

ID	答案 1	答案 2	答案 3	性别	ID	答案 1	答案 2	答案 3	性别
1	1	5	3	1	11	3	1	2	0
2	1	3	4	0	12	3	6	1	1
3	4	5	6	0	13	3	2	4	1
4	1	4	3	0	14	4	3	1	0
5	1	4	6	0	15	6	3	4	0
6	3	4	5	1	16	2	3	1	0
7	2	3	4	1	17	2	1	3	0
8	5	6	1	1	18	2	3	3	1
9	5	3	4	0	19	3	2	4	1
10	4	2	3	0	20	2	1	4	1

3.8 思考与练习

1. 打开数据文件 data3-8.sav，完成以下统计分析。

(1) 计算各科成绩的描述统计量：平均成绩、中位数、众数、标准差、方差、极差、最大值和最小值。

(2) 生成一个新变量“成绩段”，其值为各科成绩的分段：90~100 为 1，80~89 为 2，70~79 为 3，60~69 为 4，60 分以下为 5，其值标签：1—优，2—良，3—中，4—及格，5—不及格。分段以后进行频数分析，统计各分数段的人数，最后生成条形图和饼图。
2. 打开数据文件 data3-9.sav，完成以下统计分析。

(1) 对身高进行考察，分析四分位数，并生成茎叶图和箱图。

(2) 考察身高、体重和胸围的正态性。
3. 表 3.22 是对吸烟与患气管炎的调查表，试分析吸烟与患气管炎之间的关系。（用交叉列联表分析，参见数据文件：data3-10.sav。）

4. 为分析某中学学生填报志愿的倾向，设计了一道问卷调查题，每位同学可填报 3 个志愿，请按顺序依次选择打算报考的大学：

第一志愿

第二志愿

第三志愿

① 北京大学 ② 清华大学 ③ 复旦大学 ④ 中国人民大学

⑤ 北京交通大学 ⑥ 四川大学

表 3.22 吸烟人群健康状况调查表

是否吸烟	是否患气管炎	人数
是	患病	43
是	健康	162
否	患病	13
否	健康	121

问卷调查的结果存放在 SPSS 数据文件 data3-11.sav 中，按如下要求进行统计分析。

(1) 对第一、二、三志愿填报情况进行统计分析。

(2) 对各学校填报志愿的情况进行统计分析，包括人数、百分比等。

5. 打开数据文件 data3-12，按以下要求进行操作：

(1) 对变量“资讯 1”到“资讯 5”进行多重响应集的频率分析。

(2) 对性别与资讯进行交叉列联分析。

第 4 章 均值比较与 T 检验

在学习本章之前，我们需要先理解统计学中推断统计、参数估计和假设检验等相关概念。

统计学中常用的统计方法分为描述统计和推断统计两大类。描述统计仅仅针对样本数据进行处理，而推断统计则要从样本数据出发推断其总体特征。如果掌握了所研究总体的全部数据，那么只需做一些简单的统计描述，就可得到有关总体的数据特征，如方差、总体均值等，但在现实情况中，很多时候不可能或者不必对总体中的每个单位都进行测定，就需要从总体中抽取一部分单位进行测定，通过样本提供的信息来对总体信息进行推断。

利用样本数据对总体特征的推断通常有以下两种情况：（1）在总体分布已知（如总体为正态分布）的情况下，对总体包含的参数进行推断的问题称为参数检验。通常，样本量很大时，由中心极限定理可知，样本均数的抽样分布仍然是正态的，因此研究者很少去考虑参数检验的适用条件。（2）在总体分布未知的情况下，根据样本数据对总体的分布形式或特征进行推断，通常采用的统计推断方法是非参数检验方法，这部分内容我们将在第 5 章进行介绍。

对总体特征的推断一般采用参数估计和假设检验两类方式来实现。例如，我们对储户一次取钱的金额进行测定，得到了一批数据，然后求出储户一次取钱的平均金额，这就是参数估计问题。经过长期积累，知道了储户一次取钱金额的平均值和标准差，现在对某一储蓄点某一天中的取钱次数及每次取钱金额进行监控，又得到一批数据。要求该储蓄点当天一次取钱的平均金额与已知的储户一次取钱金额的平均值相比，是否有显著差异，这就是假设检验的均值比较问题。假设检验的基本思路是先对总体特征做出某种假设，然后利用样本提供的信息判断前面提出的假设是否成立，应该拒绝还是接受。参数估计和假设检验两者关系较为密切，在 SPSS 两者的结果常常会在输出窗口中同时给出。

在统计分析中，我们经常需要对两个总体的样本进行均值比较，从而推断总体存在的差异。在 SPSS 中，主要用 T 检验的方法来对两个样本进行比较，T 检验属于参数检验，要求样本来自的总体服从正态分布。T 检验的方法都包含在“分析”菜单的“比较平均值”菜单中，该菜单包括 6 个子菜单：

- ① 平均值：计算各种基本描述统计量；
- ② 单样本 T 检验：检验单个变量的均值与假设检验值之间是否存在差异；
- ③ 独立样本 T 检验：检验两组来自独立总体的样本，其独立总体的均值或中心位置是否一样；
- ④ 摘要独立样本 T 检验：用来对已知基本统计参数的样本直接进行 T 检验；
- ⑤ 成对样本 T 检验：检验两个相关的样本是否来自具有相同均值的总体；
- ⑥ 单因素 ANOVA 检验：用来推断控制变量的不同水平对观测变量是否有显著影响，这个菜单将在第 6 章中介绍。

4.1 假设检验

在做均值比较以及后续章节的各种统计分析之前，必须要理解好假设检验的原理，这里我们对其做进一步的详细介绍。假设检验也叫“显著性检验（Test of statistical significance）”，是统计学中根据一定假设条件由样本推断总体的一种方法。具体作法：先对所研究的总体做出某种假设，然后通过抽样研究，推断出应该对此假设拒绝还是接受的结论。

4.1.1 基本概念及统计原理

假设检验的基本思路是先对总体特征做出某种假设，然后利用样本提供的信息去验证前面提出的假设是否成立。如果样本数据不能充分证明和支持假设的成立，则在一定的概率条件下，应拒绝该假设；反之，如果样本数据不能充分证明和支持假设是不成立的，则不能推翻原假设。

在假设检验的过程中，涉及的几个概念如下。

1. 统计假设

要做出某些决策，常常要对总体先做出某些假设，这些假设可能正确也可能不正确，称为统计假设。它们一般是关于总体概率分布的某些陈述。

统计假设包括原假设和备择假设。

原假设：被检验的假设，通过检验可能被接受，也可能被否定。在很多情况下，我们给出一个统计假设仅仅是为了拒绝它。例如，如果要判断给定的一枚硬币是否均匀，则假设硬币是均匀的（即 $P = 0.5$ ，其中 P 是正面出现的概率）。类似地，如果要判断一种方法是否优于其他方法，则假设两种方法之间没有差异。这样的假设通常称为原假设或零假设，记为 H_0 。

备择假设： H_0 对应的假设，只有在原假设被否定后才可接受的假设，无充分理由是不能接受的。任何不同于原假设的假设都称为备择假设。例如，如果原假设是 $P = 0.5$ ，则备择假设是 $P \neq 0.5$ 。备择假设记为 H_1 。

原假设与备择假设相互排斥，并且同时只有一个正确。

拒绝域、临界点：当检验统计量取某个区域中的值时，拒绝原假设，则称该取值区域为拒绝域，称拒绝域的边界点为临界点。

2. 假设检验的两类错误

（1）第一类错误：在假设检验中拒绝了本来是正确的原假设。

H_0 本身是成立的，但通过检验却否定了它，我们称这类错误为 α 错误或第一类错误，即当原假设正确时我们却认为它错了，拒绝了原假设，也称为“弃真”错误。

（2）第二类错误：在假设检验中没有拒绝错误的原假设。

H_0 本身是不成立的，但通过检验却接受了它，那就犯了另一类错误，称为 β 错误或第二类错误，也称为“取伪”错误。

3. 显著性水平与置信水平

在做假设检验时，我们可以接受的犯第一类错误的最大概率称为检验的显著性水平，这个概率常记为 α 。显著性水平 α 对假设检验的结论有直接影响，通常抽样前就指定好，得到的结果才不会影响我们的选择。 α 为置信度或置信水平。

在实际问题中，显著性水平可以有多种选择，最普通的是 0.05 或 0.01。到底选哪个显著性水平，应根据试验的要求或试验结论的重要性而定。如果试验中对精确度的要求较高或试验结论的应用事关重大，如药物的毒性试验，则所选显著性水平应高些，即 α 值应小些。例如，如果设计一个决策法则选择的显著性水平是 0.05（5%），那么在 100 次中可能有 5 次机会使我们拒绝本该接受的假设，也就是说，我们大约有 95% 的把握做出正确的决策。此时，我们说拒绝假设的显著性水平为 0.05，即犯“拒绝本应接受的假设”这类错误的概率是 0.05。

4. 概率 P 值

P 值是当原假设正确时，观测到的样本信息出现的概率。如果这个概率很小，以至于几乎不

可能在原假设正确时出现目前的观测数据，我们就拒绝原假设。 P 值越小，拒绝原假设的理由就越充分。但怎样的 P 值才算“小”呢？通常是与预先设定的显著性水平 α 值比较，若 α 值为 0.05， P 值小于 0.05 则认为该概率值足够小，可以理解为原假设很不“显著”，应拒绝原假设。在 SPSS 软件统计结果中，“显著性”下面的值就是统计出的 P 值。

5. 单侧检验与双侧检验

(1) 双侧检验：只强调差异而不强调方向性的检验叫双侧检验。

(2) 单侧检验：强调某一方向的检验叫单侧检验。

应根据实际问题选择单侧或双侧检验。应该用单侧检验的问题若使用了双侧检验，其结果可能使结论由“显著”变为“不显著”，也可能增大 β 错误；而应该用双侧检验的问题若使用了单侧检验，则会使无方向性的问题变为错误的单方向问题。

4.1.2 小概率事件原理

在概率论中我们把发生概率小到接近于 0 的事件称为小概率事件（即在大量重复试验中出现的频率非常低）。日常生活中小概率事件是非常多的，如买彩票中大奖、雷电伤人、多胞胎等。虽然小概率事件发生的概率很小，但是一旦发生，常具有很大的影响力，因此小概率事件是不可忽视的。统计学上一般把 $P \leq 0.01$ 或 $P \leq 0.05$ 的事件（即事件发生的概率在 0.01 以下或 0.05 以下的事件）称为小概率事件，将 0.05 或 0.01 这两个域值称为小概率标准。对于某些一旦发生后果特别严重的小概率事件，其域值则需要选得更小。

设 H_0 为原假设， H_1 为与原假设对立的备择假设（对立假设），构造一个随机事件 A ，当原假设成立时随机事件 A 以很小的概率发生，该事件 A 称为小概率事件。

在统计学上，把小概率事件看成在一次特定的抽样中不可能发生的事件，称为“小概率事件实际不可能原理”。这是统计学上进行假设检验（显著性检验）的基本依据。根据这一原理，若某事件在理论上被认为在原假设成立的情况下是个小概率事件，它不会发生，而在实际中发生了，我们就推翻原来的假设，认为原假设不成立，从而接受备择假设。这个原理要在抽样理论的基础上去理解，为了更好地理解假设检验的问题，举个例子来说明：某人声称一个口袋里面的 20 个乒乓球中，有红球 10 个和白球 10 个。另一个人为了检验这种说法的正确性，便从口袋中随机抽取 10 个，发现有 2 个红球，8 个白球。请根据随机抽取结果检验口袋中是否红球和白球各 10 个。

首先，提出原假设与备择假设。从题目可知，原假设 H_0 ：口袋中红球和白球各 10 个，显然备择假设是红球和白球数不相等。其次，计算概率 P 值，即当原假设为真时，随机抽样事件发生的概率值。在本例中，如果红球和白球各 10 个，则从中抽取 2 个红球和 8 个白球的概率为 $P = C_{10}^2 C_{10}^8 / C_{20}^{10} = 0.01096$ 。最后，给定显著性水平，做出统计决策。我们给定显著性水平 $\alpha = 0.05$ ，由于 $P < 0.05$ ，我们应拒绝原假设（即认为两种颜色的球个数相等的假设不正确）。因为，如果原假设成立，本例中的抽样就是一个小概率事件。而根据小概率原理（即认为小概率事件在一次随机抽样中是不会发生的，如果发生了，则应拒绝原假设，此时拒绝原假设犯错误的概率很小，在本例中为 1.096%），我们应该拒绝原假设，即认为两种颜色的球个数不相等，而且白球比红球多。

4.1.3 假设检验的一般步骤

假设检验是对给定的总体参数值，利用样本数据对其推断，并给出接受或者拒绝的过程。假设检验依据“小概率事件实际不可能原理”，如果发生了小概率事件，我们有理由怀疑假设的正确性，从而拒绝假设检验的原假设。在具体操作中，进行假设检验时，首先应定义所谓的小概率，一般取 0.01 或 0.05，即显著性水平。显著性水平取值太小，容易发生取伪错误；取值太大，则容

易发生弃真错误。

根据以上假设检验原理及小概率事件的讨论，可归纳出参数检验的基本步骤如下。

第 1 步 给出检验问题的原假设、备择假设。

根据检验问题的要求，将需要检验的最终结果作为原假设。例如，需要检验某学校的高考数学平均成绩是否同往年的平均成绩一样，都为 75，由此可做出原假设 $H_0:\mu=75$ 。

第 2 步 选择检验统计量。

在统计推断中，总是通过构造样本的统计量并计算统计量的概率值进行推断，一般构造的统计量应服从或近似服从常用的已知分布，例如均值检验中最常用的 T 分布和 F 分布等。

第 3 步 规定显著性水平。

这里的显著性水平指的是当假设正确时被拒绝的概率，即弃真概率，一般取 0.01 或 0.05。

第 4 步 计算检验统计量的观测值及其发生的概率值。

在给定原假设前提下，计算统计量的观测值和相应概率 P 值。概率 P 值就是在原假设 H_0 成立时检验统计量的观测值发生的概率，该概率值间接地给出了样本值在原假设成立前提下出现的概率，对此可以依据一定的标准来判断其发生的概率是否为小概率。

第 5 步 在给定显著性水平条件下，做出统计推断结果。

当检验统计量的概率 P 值小于显著性水平时，则认为拒绝原假设而犯弃真错误的概率小于显著性水平，即低于预先给定的水平，也就是说，犯错误的概率小到我们能容忍的范围，这时可以拒绝原假设；反之，当检验统计量的概率 P 值大于显著性水平时，则认为拒绝原假设而犯弃真错误的概率大于预先给定的容忍水平，这时不应该拒绝原假设。

所以，在 SPSS 的检验问题中，都是利用概率 P 值和显著性水平进行比较，做出拒绝或者接受原假设结论的。SPSS 中系统自动计算概率 P 值，但显著性水平应该由用户事先设定。

4.2 平均值分析

4.2.1 平均值分析的概念及统计原理

与第 3 章中的“描述统计”菜单中计算某一样本总体均值相比，平均值分析可以对样本进行分组计算，比较指定变量的描述性统计量包括均值、标准差、总和、观测量数、方差等一系列单变量描述性统计量，还可以给出方差分析表和线性检验结果。如果分组变量为多个，还应指定这些分组变量之间的层次关系。

均值过程中系统默认的描述统计量可按分组给出指定变量的均值、标准差、观测量数等，对话框中的选项可以给出其他更加丰富的描述统计量。

平均值的计算公式：
$$\bar{x}_1 = \frac{\sum_{i=1}^n x_{1i}}{n}$$

4.2.2 平均值 SPSS 实例分析

【例 4-1】表 4.1 是各地区分性别受教育程度的人口数量，利用平均值分析比较受教育程度是否受性别的影响。（参见数据文件：data4-1.sav。）

第 1 步 数据组织

根据表 4.1 生成 SPSS 数据文件，建 3 个变量：“性别”、“教育”、“人口数量”，建立的数据文件存入文件 data 4-1.sav 中。

表 4.1 各地区分性别受教育程度的人口数量

地 区	小学		初中		高中	
	男	女	男	女	男	女
北 京	885	1002	2138	1920	1550	1638
天 津	924	1060	1803	1583	1076	1042
河 北	8563	9960	15323	12862	3846	2954
山 西	3785	4161	7287	6572	2394	1885
内蒙古	3060	3265	4617	3736	1722	1371

第2步 平均值分析设置

(1) 选择菜单“分析→比较平均值→平均值”，打开“平均值”对话框，按图 4-1 所示进行设置。



图 4-1 “平均值”对话框

➤ 候选变量框：列出数据文件中所有的变量。

➤ 因变量列表：从左侧的变量列表中选择待分析的变量，可选择一个或多个。

➤ 自变量列表：从左侧的变量列表中选择分组变量，可选择一个或多个。还可单击“下一个”定义多层分组变量，每层分组变量中也可以有多个变量。

(2) “平均值分析：选项”对话框设置。单击图 4-1 对话框中“选项(O)”按钮，弹出如图 4-2 所示的“平均值：选项”子对话框。该对话框由如下几部分组成。

➤ “统计(S)”选项框：在该文本框中列出可以选择的描述性统计量，这些统计量的具体含义同描述性统计分析中的统计量含义一样，此处不再详细描述。

➤ “单元格统计(C)”框：列出要输出的统计量。默认输出平均值、个案数和标准差。

➤ “第一层的统计”选项组：该选项组定义是否进行分组第一层变量的方差分析(Anova 表和 Eta)和线性相关度检验。

第3步 主要结果及分析。

完成以上操作步骤后，单击图 4-1 上的“确定”按钮，在输出文件中，得到均值分析的结果，运行结果如表 4.2~4.5 所示，具体分析如下：

(1) 表 4.2 给出了样本的数据摘要，从表中可以看出，30 组数据全部有效。

(2) 表 4.3 显示的是按性别分组的受教育（从小学到高中的教育）的人口数量的基本信息，从表中可以看出，不同性别受教育的人口数量其均值和标准差都比较接近。



图 4-2 “平均值：选项”对话框

表 4.2 个案处理摘要表

个案处理摘要						
	个案					
	包括		排除		总计	
	个案数	百分比	个案数	百分比	个案数	百分比
人口数量 * 性别	30	100.0%	0	0.0%	30	100.0%

表 4.3 按性别分组的均值报告

报告			
人口数量			
性别	平均值	个案数	标准差
男	3931.53	15	3881.083
女	3667.40	15	3528.010
总计	3799.47	30	3646.720

(3) 表 4.4 是性别的单因素方差分析，在第 6 章会详细介绍方差分析，此处不再详细讲述。表中的显著性概率 P 值远大于 0.05，说明不同性别受教育的人口数量没有显著性差异。

表 4.4 第一层变量的方差分析

ANOVA 表 a

		平方和	自由度	均方	F	显著性
人口数量 * 性别	组间	(组合) 523248.133	1	523248.133	.038	.847
	组内	385135243.333	28	13754830.119		
	总计	385658491.467	29			

a. 分组变量 性别 是字符串，因此无法计算线性相关度检验。

(4) 表 4.5 是人口数量与性别的相关性度量表。此时的 Eta 和 Eta 平方取值都很小,说明性别和受教育的人口数量的相关性很差,这也和单因素方差分析表的结论是一致的。

以上的分析结果是按性别分组的分组结果,在 4.2.2 节中我们讲到平均值分析中可以选择分层变量对变量进行分层分析,如果我们在图 4-1 的对话框中单击“下一个”将“受教育程度”选为第二层分组变量,分析结果还会按照受教育程度进行分类,输出结果如表 4.6 所示,读者可自行分析。

表 4.5 相关性测量表

相关性测量		
	Eta	Eta 平方
人口数量 * 性别	.037	.001

表 4.6 按“性别*受教育程度”分组的均值报告

报告				
人口数量				
性别	受教育程度	平均值	个案数	标准差
男	初 中	6233.60	5	5539.503
	高 中	2117.60	5	1075.565
	小 学	3443.40	5	3137.147
	总计	3931.53	15	3881.083
女	初 中	5334.60	5	4649.796
	高 中	1778.00	5	728.246
	小 学	3889.60	5	3662.595
	总计	3667.40	15	3528.010
总计	初 中	5784.10	10	4844.783
	高 中	1947.80	10	884.248
	小 学	3666.50	10	3223.575
	总计	3799.47	30	3646.720

4.3 单样本 T 检验

4.3.1 基本概念及统计原理

1. 单样本 T 检验的概念

单样本 T 检验利用来自某总体的样本数据,推断该总体的均值与指定的检验值之间是否存在显著性差异,它是对总体均值的假设检验。为此,给出检验均值 μ_0 , 原假设: $\mu = \mu_0$, 其中 μ 为总体均值,即认为总体均值与检验值 μ_0 之间无显著性差异。

例如,从新生入学成绩的抽样数据推断平均成绩是否为 75 分;在人口普查中,某地区职工今年的平均收入是否和往年的平均收入有显著差异。

单样本 T 检验涉及的是单个总体,并采用 T 检验的方法,因此称为单样本 T 检验。

单样本 T 检验的前提是样本来自的总体应服从或近似服从正态分布,如果总体不是来自正态分布或不清楚总体的分布情况,则不能用单样本 T 检验。

2. 单样本 T 检验的检验统计量

单样本 T 检验的前提是总体服从正态分布 $N(\mu,\sigma^2)$, 其中 μ 为总体均值, σ^2 为总体方差。

如果样本容量为 n ，样本均值为 \bar{X} ，则 \bar{X} 仍服从正态分布，即 $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ 。

在原假设成立的条件下，均值检验使用 t 统计量，构造的 t 统计量：

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

式中， μ 用 μ_0 代入， t 统计量服从自由度为 $n-1$ 的 T 分布， S 为样本标准差。

SPSS 的操作结果中还显示均值标准误差 (Std.Error Mean)，计算公式： $\frac{S}{\sqrt{n}}$ ，即统计量的分母部分。

在给定原假设的前提下，SPSS 将检验值 μ_0 代入 t 统计量，得到检验统计量观测值，以及根据 T 分布的分布函数计算出的概率 P 值。

3. 单样本 T 检验的步骤

在给定样本来自正态总体的假设下，单样本 T 检验作为假设检验的一种方法，其基本步骤与假设检验的步骤一样。

4.3.2 单样本 T 检验 SPSS 实例分析

【例 4-2】某生产食盐的生产线，其生产的袋装食盐的标准质量为 500g，现随机抽取 10 袋，其质量分别为 495g, 502g, 510g, 497g, 506g, 498g, 503g, 492g, 504g, 501g。假设数据呈正态分布，请检验生产线的工作情况。（参见数据文件：data 4-2.sav。）

这是一个典型的比较样本均值和总体均值的 T 检验问题，进行单样本 T 检验的具体步骤如下：

第 1 步 数据组织。

首先建立 SPSS 数据文件，只需建立一个变量“Weight”，录入相应的数据即可，建立的数据文件存入文件 data4-2.sav 中。

第 2 步 单样本 T 检验分析设置。

(1) 选择菜单：“分析→比较平均值→单样本 T 检验 (S)”，打开“单样本 T 检验”主对话框，确定要进行 T 检验的变量并输入检验值，按图 4-3 所示进行设置。



图 4-3 “单样本 T 检验”对话框

① 候选变量框：对话框中左边的列表框为候选变量框，列出了数据文件中所有可以进行 T 检验的变量。

② 检验变量：用来存放要进行 T 检验的变量。从候选变量框中选择需要变量并移入此框中，可同时选择多个变量，此时，SPSS 将分别产生多个变量的 T 检验分析结果。

③ 检验值：输入待检验的值，用来检验产

生的样本均值与检验值有无显著性差异。

(2) “单样本 T 检验：选项”对话框设置：指定置信水平和缺失值的处理方法。

在图 4-3 的对话框中单击“选项 (O)…”按钮，打开“单样本 T 检验：选项”对话框，并按图 4-4 所示设置本次检验的置信水平并选择对缺失值的处理方式。该对话框主要包含以下几部分。

① 置信区间百分比：设置样本均值与总体均值之差的置信区间，该文本框中可以输入在 1 ~

99 之间的任意值，一般取为 90、95、99 等数值，系统默认值为 95。

② 缺失值：在该选项组中可以选择缺失值的处置方式，包括以下两个选项。

➤ 按具体分析排除个案 (A)：在检验过程中，仅剔除参与分析的缺失值。

➤ 成列排除个案 (L)：剔除所有含有缺失值的个案。

本例中设置置信水平为 95%，为了保留更多数据样本，缺失值的处理方式选择“按具体分析排除个案”。

第 3 步 主要结果及分析。

完成以上操作步骤后，单击图 4-3 中的“确定”按钮，运行结果如表 4.7 和表 4.8 所示，具体分析如下。

(1) 表 4.7 给出了单样本 T 检验的描述性统计量，包括个案数、平均值、标准差、标准误差平均值。

(2) 表 4.8 是单样本 T 检验结果表，用于比较的检验值为 500，并依次给出了检验统计量 (t)、自由度 (df)、双尾检测概率 P 值、样本均值与检验值的差 (均值差值)、均值差的 95% 置信区间。

当置信水平为 95% 时，显著性水平为 0.05，从表 4.8 中可以看出，双尾检测概率 P 值为 0.650，大于 0.05，故原假设成立，也就是说，抽样袋装食盐的质量与 500 克无显著性差异，有理由相信生产线工作状态正常。

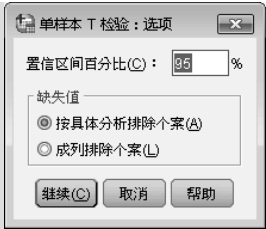


图 4-4 “单样本 T 检验：选项”对话框

表 4.7 单样本统计量

单样本统计				
	个案数	平均值	标准差	标准误差平均值
weight	10	500.8000	5.39135	1.70489

表 4.8 单样本 T 检验结果表

单样本检验						
	检验值 = 500					
	t	自由度	显著性 (双尾)	平均值差值	差值 95% 置信区间	
					下限	上限
weight	.469	9	.650	.80000	-3.0567	4.6567

4.4 独立样本 T 检验

4.4.1 基本概念及统计原理

1. 独立样本 T 检验的概念

单样本 T 检验用来检验样本均值和总体均值是否有显著性差异，而独立样本 T 检验利用来自某两个总体的独立样本，推断两个总体的均值是否存在显著差异。因此，其原假设 H_0 为 $\mu_1 = \mu_2$ ，即假设两总体均值相等，备择假设为 $\mu_1 \neq \mu_2$ ，即假设两总体均值不等。

例如，为比较两种牧草对奶牛的饲养效果，随机从奶牛群中选取喂养不同牧草的奶牛各 10

头，记录每日平均产奶量，根据记录的数据推断两种牧草对奶牛饲养效果有无显著性差异。

独立样本 T 检验采用 T 检验的方法，涉及两个总体，要求两组样本相互独立，即从一总体中抽取的一组样本对从另一总体中抽取的一组样本没有任何影响，两组样本的个案数目可以不等。因此，独立样本 T 检验的前提是样本来自的总体应服从或近似服从正态分布，且两组样本相互独立。

2. 独立样本 T 检验的检验统计量

独立样本 T 检验的前提是两个独立的总体分别服从 $N(\mu_x, \sigma_x^2)$ 和 $N(\mu_y, \sigma_y^2)$ 。在原假设成立的条件下，独立样本 T 检验使用 t 统计量。构造独立样本 T 检验的 t 统计量分为两种情况。

(1) 当样本方差相等时， t 统计量定义为

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_\omega \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.1}$$

式中， n_1 和 n_2 分别为两样本容量； $S_\omega^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ ， S_1 和 S_2 分别为两样本标准差。该统计量服从自由度为 $n_1 + n_2 - 2$ 的 T 分布。

(2) 当样本方差不等时， t 统计量定义为

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{4.2}$$

可见，独立样本 T 检验的结论在很大程度上取决于两个总体的方差是否相等。这就要求在检验两总体均值是否相等之前，首先应对两总体方差是否相等进行检验，称为方差齐性检验。

SPSS 中利用 Levene F 方差齐性检验方法检验两总体方差是否存在显著差异。进行 Levene F 方差齐性检验，首先提出原假设 $H_0: \sigma_1^2 = \sigma_2^2$ ；执行检验过程中，若概率 P 值小于给定的显著性水平（一般为 0.05），则拒绝原假设，认为两个总体的方差不等；否则认为两个总体的方差无显著性差异。

在给定原假设的前提下，SPSS 将检验值 0 代入 t 统计量的 $\mu_1 - \mu_2$ 部分，得到检验统计量观测值，并根据 T 分布的分布函数计算出概率 P 值。

3. 独立样本 T 检验的一般步骤

在两样本来自正态总体且相互独立的假设下，独立样本 T 检验作为假设检验的一种方法，其基本步骤与假设检验的步骤是一样的。

4.4.2 独立样本 T 检验 SPSS 实例分析

【例 4-3】 为比较两种不同品种玉米的产量，分别统计了 8 个地区的单位面积产量，具体数据如表 4.9 所示。假定样本服从正态分布，且两组样本相互独立，试比较在置信度为 95% 的情况下，两种玉米产量是否有显著性差异。（参见数据文件：data4-3.sav。）

表 4.9 两品种玉米单位面积产量

品种 A	85	87	56	90	84	94	75	79
品种 B	80	79	58	90	77	82	75	65

进行独立样本 T 检验的具体步骤如下。

第 1 步 数据组织。

根据表 4.9，在 SPSS 数据文件中建立两个变量，分别为“品种”、“产量”，度量标准分别为“名义”、“度量”，变量“品种”的值标签：a—品种 A，b—品种 B，录入数据后，保存名为 data4-3.sav 的 SPSS 数据文件。

第 2 步 独立样本 T 检验设置。

(1) 菜单选择：“分析→比较平均值→独立样本 T 检验”，打开“独立样本 T 检验”对话框，确定要进行 T 检验的变量，确定分组变量。按图 4-5 所示进行设置。

该对话框主要由以下几部分组成。

- ① 候选变量框：图 4-5 左侧变量列表框，列出数据文件中可以进行 T 检验的变量。
- ② 检验变量 (T)：从候选变量框中选择要进行 T 检验的变量移入此框中，可同时选择多个变量，此时，SPSS 将分别产生多个变量的 T 检验分析结果。
- ③ 分组变量 (G)：选择分组变量，在选择变量进入“分组变量”框后，“定义组 (D)…”按钮将被激活。
- ④ “定义组 (D)”按钮：定义变量的分组方法。单击该按钮，弹出如图 4-6 所示的对话框。



图 4-5 “独立样本 T 检验”对话框



图 4-6 “定义组”对话框

用特定的变量值分组，当变量的取值等于“组 1 (1)”文本框中的值时，将其划为第 1 组；当变量的取值等于“组 2 (2)”文本框中的值时，将其划为第 2 组。

(2) “选项”对话框设置：指定置信水平和缺失值的处理方法。

在图 4-5 的对话框中单击“选项 (O)…”按钮，打开“独立样本 T 检验：选项”对话框，具体选项内容及设置与单样本 T 检验相同。

第 3 步 运行结果及分析。

完成以上操作步骤后，单击图 4-5 中的“确定”按钮，运行结果如表 4.10 和表 4.11 所示，具体意义分析如下。

(1) 表 4.10 是本例独立样本 T 检验按组统计的基本描述统计量，包括两个样本的平均值、标准差和标准误差平均值。

表 4.10 独立样本 T 检验按组统计的基本描述统计量

组统计					
	玉米品种	个案数	平均值	标准差	标准误差平均值
单位面积产量	品种 A	8	81.2500	11.80496	4.17368
	品种 B	8	75.7500	10.02497	3.54436

(2) 表 4.11 是独立样本 T 检验的均值检验结果。表中给出了两种 T 检验的结果，分别为在样本方差相等情况下的一般 T 检验结果和在样本方差不等情况下的校正 T 检验结果。数理统计学中检查不同样本的总体方差是否相等即方差齐性检验 (Homogeneity of variance test) 通常采用 Hartley 检验、Bartlett 检验或 Levene 检验方法。两种 T 检验结果到底应该选择哪一个？这就取决于表 4.11 中的“莱文方差等同性检验”一项，即用莱文 (levene) 方法检验的方差齐性结果。对于齐性，这里采用的是 F 检验，表中第二列是 F 统计量的值，为 0.104，第三列是对应的概率 P 值，为 0.752。如果显著性水平为 0.05，由于概率 P 值大于 0.05，可以认为两个总体的方差无显著性差异，即方差具有齐性。

在方差具有齐性的情况下，独立样本 T 检验的结果应该看表中的“假定等方差”一行，第 5 列为相应的双尾检测概率 (显著性概率 P 值 (双尾)) 0.332，在显著性水平为 0.05 的情况下， T 统计量的概率 P 值大于 0.05，故不应拒绝原假设，因此认为两样本的均值是相等的，在本例中，不能认为两种玉米品种的产量有显著性差异。

表 4.11 独立样本 T 检验结果表

独立样本检验										
		莱文方差等同性检验		平均值等同性 t 检验						
		F	显著性	t	自由度	显著性 (双尾)	平均值差值	标准误差差值	差值 95% 置信区间	
									下限	上限
单位面积产量	假定等方差	.104	.752	1.004	14	.332	5.50000	5.47560	-6.24398	17.24398
	不假定等方差			1.004	13.642	.333	5.50000	5.47560	-6.27297	17.27297

4.4.3 摘要独立样本 T 检验

SPSS 23 中，在“分析→比较平均值”菜单下新增了一个子菜单“摘要独立样本 T 检验”，在该菜单中，用户可以直接填写统计样本的个案数、平均值、标准差等统计参数值，SPSS 会根据摘要数据计算 T 检验。如图 4-7 所示，我们可以将表 4.10 的基本描述统计量填写到对话框，直接计算两个玉米品种的 T 检验，输出结果表 4.12 和表 4.11 是一致的，但表 4.12 中方差齐性采用的是 Hartley 方法，该方法适用于样本量相等的场合，得到的显著性为 0.3082，大于显著性水平 0.05，同样，方差具备齐性。



图 4-7 “根据摘要数据计算 T 检验”对话框

表 4.12 摘要独立样本 T 检验结果表

独立样本检验					
	平均值差值	标准误差差值	t	自由度	显著性（双尾）
假定等方差	5.500	5.142	1.070	16.000	.301
不假定等方差	5.500	5.241	1.049	13.827	.312
Hartley 等方差检验: F = 1.387, 显著性 = 0.3082					

4.5 配对样本 T 检验

4.5.1 基本概念及统计原理

1. 配对样本 T 检验的概念

配对样本 T 检验用于检验两组相关样本是否来自相同均值的正态总体，即推断两个总体的均值是否存在显著差异。其原假设为 $H_0: \mu_1 - \mu_2 = 0$ ，其中， μ_1 和 μ_2 分别为第一个总体和第二个总体的均值。

配对的概念是指两组样本的各样本值之间存在着对应关系，配对样本的两组样本值之间的配对是一一对应的，并且两组样本的容量相同。配对样本 T 检验与独立样本 T 检验的差别之一是要求样本是配对的。所谓配对样本，可以是个案在“前”“后”两种状态下某属性的两种状态，也可以是对某事物两个不同侧面或方面的描述。其差别在于抽样不是相互独立的，而是相互关联的。

例如，考察同一组人在参加一年的长跑锻炼前后的心率是否有显著差异。这时，每个人一年前的心率和一年后的心率是相关的。再如，为研究某种减肥茶是否有显著的减肥效果，需要对肥胖人群喝茶前与喝茶后的体重进行分析，通常采用配对的抽样方式：首先从肥胖人群中随机抽取部分人，记录他们喝茶前的体重，喝茶一段时间后重新测量这些肥胖人群的体重，这样获得的两组样本均是配对样本。

2. 配对样本 T 检验的数学思想

配对样本 T 检验须求出每对观测值之差，所有样本值的观测值之差形成一个新的单样本，显然，如果两个样本的均值没有显著差异，则样本值之差的均值应该接近零，这实际上转换成了一个单样本的 T 检验。所以，配对样本 T 检验就是检验差值所来自的总体其均值是否为零，这就要求差值来自的总体服从正态分布。

3. 配对样本 T 检验的检验统计量

在配对样本 T 检验中，设 x_{1i} ， x_{2i} ($i = 1, \cdots, n$) 分别为配对样本。其样本差值 $d_i = x_{1i} - x_{2i}$ ，此时检验统计量

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)}{S / \sqrt{n}} \tag{4.3}$$

式中， \bar{d} 为 d_i 的均值； S 为 d_i 的标准差； n 为样本数。当 $\mu_1 - \mu_2 = 0$ 时， t 统计量服从自由度为 $n-1$ 的 T 分布。

SPSS 中将计算两组样本的差值，并将相应数据代入上式的 T 检验统计量计算公式中，计算出 T 统计量的观测值和对应的概率 P 值。

4. 配对样本 T 检验主要步骤

配对样本 T 检验作为假设检验的一种方法，其基本步骤与假设检验的步骤是一样的。

4.5.2 配对样本 T 检验 SPSS 实例分析

【例 4-4】 以下是某大学 15 位跆拳道选手平衡训练的数据，检验实验前、后平衡训练成绩是否有差异。（参见数据文件：data4-4.sav。）

训练前：86，77，59，79，90，68，85，94，66，72，75，72，69，85，88
训练后：78，81，76，92，88，76，93，87，62，84，87，95，88，87，80
对该数据文件中的两个变量进行配对样本 T 检验的具体步骤如下。

第 1 步 数据组织。

首先建立 SPSS 数据文件，建立两个变量：“训练前”、“训练后”，录入相应数据，保存为文件 data4-4.sav。

第 2 步 配对样本 T 检验设置。

（1）选择菜单：“分析→比较平均值→成对样本 T 检验”，弹出“成对样本 T 检验”对话框，确定要配对分析的变量，按图 4-8 所示进行设置。



图 4-8 “成对样本 T 检验”对话框

该对话框主要由以下几部分组成。

- 候选变量框：图 4-8 左侧变量列表框，列出数据文件中可以进行配对样本 T 检验的变量。
 - 配对变量（V）：该列表框中的变量作为分析变量，总是成对出现，可以有多对分析变量。
- （2）“选项”对话框设置：指定置信水平和缺失值的处理方法。

关于该对话框的各选项意义及其设置，在前面已有讲述，可以参考前文。

第 3 步 运行结果及分析。

完成以上操作步骤后，单击图 4-8 中的“确定”按钮，运行结果如表 4.13～表 4.15 所示，具体意义分析如下。

（1）表 4.13 是成对样本的基本描述性统计量，包括每一组样本的均值、样本容量、标准差和均值标准误。从两对样本的均值变化可以看出，均值都有一定量的变化，但是否存在显著性差异，还必须通过计算相应的 t 统计量来确定。

（2）表 4.14 是成对样本 T 检验的简单相关关系检验结果。表中第 3 列为训练前和训练后样本的相关系数，第 4 列是相关系数的检验 P 值。

表 4.13 配对样本 T 检验的基本描述统计量

配对样本统计					
		平均值	个案数	标准差	标准误差平均值
配对 1	训练前	77.67	15	10.104	2.609
	训练后	83.60	15	8.433	2.177

表 4.14 配对样本相关性检验

配对样本相关性				
		个案数	相关性	显著性
配对 1	训练前 & 训练后	15	.407	.132

从表中可以看出，在显著性水平为 0.05 时，概率 P 值为 0.132，大于 0.05，接受原假设，可以认为训练前后的成绩没有明显的线性关系。

(3) 表 4.15 是配对样本 T 检验的最终结果。训练前、后配对样本的平均差值为 5.933，差值的标准差为 10.187，差值的均值标准误为 2.630，置信度为 95%时差值的置信下限和置信上限共同构成了该差值的置信区间 (11.575, 0.292)，统计量的观测值 t 为-2.256，自由度 df 为 14，显著性概率 P 值（双侧）为双尾检验概率 P 值，在显著性水平为 0.05 时，由于概率 P 值为 0.041，小于 0.05，拒绝原假设，即认为 $u_1-u_2 \neq 0$ ，故可以认为训练对成绩有显著效果。

表 4.15 成对样本 T 检验结果

配对样本检验									
		配对差值					t	自由度	显著性 (双尾)
		平均值	标准差	标准误差 平均值	差值 95% 置信区间				
					下限	上限			
配对 1	训练前 - 训练后	-5.933	10.187	2.630	-11.575	-.292	-2.256	14	.041

4.6 典型 案 例

4.6.1 蛋白饲料对小白鼠体重影响分析

为了研究动物体重受蛋白饲料的影响情况，采用完全随机设计的方法，将 19 只体重、出生日期等相仿的小白鼠随机分为两组，其中一组喂养高蛋白饲料，另一组喂养低蛋白饲料，然后观察喂养 8 周后各小白鼠所增体重 (mg) 情况，数据中含有 19 个观测样本，代表了 19 个小白鼠，19 个小白鼠分为 2 组。有 2 个属性变量：group（分组）、weight（体重），具体数据如表 4.16 所示。（数据来源：宇传华 等，《SPSS 与统计分析》，电子工业出版社；参见数据文件：data4-5.sav。）

案例分析：因为研究的是两组喂养不同饲料的小白鼠的体重，所以两组数据应该是独立的，在数据组织时，每个个案的存放位置可以随意变动，对分析结果没有影响，并且两组的个案数不相同，所以可先用探索分析对两组数据进行正态性检验，得到两组数据都是正态分布的，因此采用独立样本 T 检验；如果两组数据不是正态分布的，应采用第 5 章所介绍的独立样本非参数检验来分析两组数据所来自两个总体的均值差异的显著性。

表 4.16 小白鼠体重

Group (分组)	Weight (体重)	Group (分组)	Weight (体重)
1	134	2	70
1	146	2	118
1	104	2	101
1	119	2	85
1	124	2	107
1	161	2	132
1	107	2	94
1	83	2	97
1	113	2	123
1	129		

4.6.2 健康教育对儿童血红蛋白水平的影响分析

某地区随机抽取 12 名贫血儿童的家庭，实行健康教育干预三个月，干预前后儿童的血红蛋白（%）测量结果如表 4.17 所示，试问干预前后该地区贫血儿童血红蛋白（%）平均水平有无变化？（数据来源：武松 等，《SPSS 统计分析大全》，清华大学出版社；参见数据文件 data4-6.sav。）

案例分析：两个样本值之间是一一配对的，每个配对是在一名儿童“前”“后”两种状态下的两种结果，并且两个样本的容量相同，因此应该用配对样本 T 检验。

表 4.17 儿童血红蛋白测量结果

序号	干预前	干预后	序号	干预前	干预后
1	36	45	7	42	70
2	46	64	8	45	45
3	53	66	9	25	50
4	57	57	10	55	80
5	65	70	11	51	60
6	60	55	12	59	60

4.6.3 储户的储蓄金额的差异分析

某银行调查了 400 名储户的性别、年龄、受教育年限、储蓄金额等信息，部分数据见表 4.18，请分析储蓄金额在各类别的储户上是否有差异？（数据来源：李昕，等，《SPSS 22.0 统计分析——从入门到精通》，电子工业出版社；参见数据文件 data4-7.sav。）

表 4.18 储户信息调查表

性别	年龄	年龄段	受教育年限	储蓄金额
男	79	65 +	12	305100
男	32	< 35	17	111875
女	50	45 - 64	6	135600
女	56	45 - 64	8	149160
女	51	45 - 64	17	237300
男	48	45 - 64	12	152550
女	29	< 35	13	211875
女	40	35 - 44	13	110175
...

案例分析：很明显，本题是对样本进行分组均值比较，题中给出了性别、年龄段、受教育年限三个分组字段（三个自变量），读者可以将这三个字段都选到第一层自变量列表，也可以分层进行更深入的分析：如性别为第一层，年龄段为第二层。注意，如果对受教育年限进行分组比较，最好对其进行离散化处理，然后创建新的分组变量。

4.7 思考与练习

- 1. 什么是“弃真”错误？怎样判断接受还是拒绝原假设？
- 2. 进行参数检验的前提条件是什么？
- 3. 表 4.19 是某班学生的高考数学成绩，试分析该班的数学成绩与全国的平均成绩 70 分之间是否有显著性差异。（参见数据文件：data4-8.sav。）

表 4.19 某班学生数学成绩

序号	成绩	序号	成绩	序号	成绩
1	63	10	94	19	70
2	99	11	98	20	65
3	81	12	73	21	84
4	77	13	89	22	84
5	68	14	98	23	95
6	79	15	77	24	61
7	80	16	67	25	69
8	63	17	69	26	73
9	87	18	81	27	60

- 4. 在某次测试中，随机抽取男女学生的成绩各 10 名，数据如下。
男：99 79 59 89 79 89 99 82 80 85
女：88 54 56 23 75 65 73 50 80 65
假设样本总体服从正态分布，比较置信度为 95%的情况下男女得分是否有显著性差异。（参见数据文件：data4-9.sav。）

- 5. 某医疗机构为研究某种减肥药的疗效，对 16 位肥胖者进行为期半年的观察测试，测试指标为使用该药之前和之后的体重，数据如表 4.20 所示。假设体重近似服从正态分布，试分析服药前后，体重是否有显著变化。（参见数据文件：data4-10.sav。）

表 4.20 服药前后的体重变化

	体 重
服药前	198 237 233 179 219 169 222 167 199 233 179 158 157 216 257 151
服药后	192 225 226 172 214 161 210 161 193 226 173 154 143 206 249 140

- 6. 试分析表 4.18 中 400 名储户的平均年龄是否为 50 岁？大于等于 50 岁和小余 50 岁的储户的储蓄金额是否存在显著差异？（参见数据文件 data4-7.sav。）
- 7. 某班两个学期的英语期末成绩如表 4.21（部分数据）所示，请分析这个班的同学两个学期的英语成绩是否存在显著差异？（数据来源：冯岩松 等，《SPSS 22.0 统计分析应用教程》，清华大学出版社；参见数据文件：data4-11.sav。）

表 4.21 某班两个学期的英语期末成绩表

学号	第一学期英语成绩	第二学期英语成绩
1	62	72
2	63	65
3	66	54
4	67	71
5	69	67
6	73	80
7	74	74
8	78	78
...

第 5 章 非参数检验

在数据分析过程中，由于种种原因，经常无法获知总体分布形态，也很难对总体分布形态做出较为准确的假定，但又希望能根据样本数据对总体的分布形式或特征进行推断，获得尽可能多的总体信息，参数检验的方法不再适用，此时通常采用非参数检验的方法。非参数检验是在总体分布未知的情况下，利用样本数据对总体分布形态等进行推断的方法，在推断过程中可不涉及有关总体分布的参数，而是检验总体某些有关的性质，如总体的分布位置、分布形状之间的比较等。如以均值比较为例，参数检验比较的是各样本的均值是否相等，而非参数检验比较的是各样本的中位数（中位数是分布位置的一种衡量）是否相等。

与参数检验的原理相同，非参数检验过程也是先根据问题提出原假设，然后利用统计学原理构造出适当的统计量，最后利用样本数据计算统计量的概率 P 值，与显著性水平 α 进行比较，得出拒绝或者接受原假设的结论。

SPSS 23 中进行非参数检验可由“分析”菜单中“非参数检验”菜单中的三个三级子菜单：单样本（O）、独立样本（I）、相关样本（R）来实现，在另一个三级子菜单“旧对话框”中保留了 SPSS 18 之前的低版本的菜单供用户使用，包括卡方（C）、二项（B）、游程（R）、单样本 K-S（1）、2 个独立样本、 K 个独立样本（K）、2 个相关样本（L）、 K 个相关样本（S）等共 8 个四级子菜单，这 8 个四级子菜单也都包含在了前面新的三个三级子菜单中。8 个菜单中前四种方法通常用来做分布的拟合优度检验，即检验样本所在的总体是否服从某个已知的理论分布。后四种方法通常用于分布位置检验，即检验样本所在总体的分布位置或形状是否相同。

5.1 参数检验与非参数检验的比较

1. 参数检验和非参数检验的区别

参数检验和非参数检验最本质的区别：参数检验需要事先确定或假定总体的分布，非参数检验则不需要假定总体的分布，而是直接用样本来推断总体的分布。

可以通过是否假定总体的分布来区分参数检验和非参数检验，除此之外，二者之间还可以从很多方面来区分。

（1）研究的对象和目标不同。参数检验研究的是总体的参数，不涉及总体的分布检验，一旦总体的参数确定，总体的分布也就确定了；非参数检验的目标是直接从样本推导总体的分布或两个总体的分布是否相同。

（2）研究的统计量有所不同。参数检验中很少用到秩来构造统计量，无论样本量大小都能对总体进行推断；非参数检验中常用秩、秩和等来构造统计量，且常要求样本量较大。

2. 非参数检验的优点

与参数检验相比，非参数检验具有以下几个优点：

（1）它对总体分布一般不做过多的限制性假设，任何分布都可以用非参数检验进行研究，从应用范围看，其应用范围大于参数检验。

(2) 由于非参数检验不依赖于总体的分布形式, 因而它天然具有稳健性特征。

(3) 对资料的测量水平要求不高, 这给资料的搜集带来了很大的方便, 可以大大减轻统计资料的搜集工作量。同时, 也为属性资料研究提供了广泛的基础。

(4) 非参数检验比较直观, 很容易理解, 不需要太多数学知识和统计理论。

多数非参数检验的运算比较简单, 可以较快地取得统计结果。

非参数检验的上述优点表明, 在实际问题的研究中, 它是一种比较有用的统计方法。

3. 非参数检验的缺点

有些人主张用非参数检验取代参数检验, 这种看法有点偏激, 因为非参数检验毕竟存在着一些自身难以克服的不足, 表现在:

(1) 两者的效率有差距。非参数检验主要处理定序资料, 这类资料的测量尺度比较低, 如果把那些能够用参数检验处理的资料转化为定类和定序资料, 必然会丢失检验数据的一部分信息, 因此非参数检验的有效性或检验效率不如参数检验。

(2) 当样本容量比较大时, 非参数检验的计算也比较繁杂、困难。

(3) 参数检验与非参数检验各有各的特点, 并非所有的参数检验都能转用非参数检验。

总之, 参数检验和非参数检验应该结合起来使用, 做到互相补充。如果条件允许, 最好使用参数检验。

5.2 单样本的非参数检验

5.2.1 基本概念及设置

单样本非参数检验使用一个或多个非参数检验方法来识别单个总体的分布情况, 不需要待检验的数据呈正态分布。

SPSS 的单样本非参数检验方法包括卡方检验、二项分布检验、游程检验、K-S 检验及 Wilcoxon 符号检验五种。SPSS 23 提供了两种方法进行单样本的非参数检验, 一种是“旧对话框”, 提供了 SPSS 18 以前的界面供老用户使用; 另一种是新版本中新增的“单样本(O)”菜单, 将卡方检验、二项分布检验、游程检验等单样本的非参数检验集中在一起, 各检验的设置相较于低版本有不小的变化。

在 SPSS 23 中, 单样本非参数检验的对话框有三个选项卡, 分别为“目标”、“字段”和“设置”。所有单样本的非参数检验有一些共同的设置, 我们以例 5-1 的前部分操作为共同设置的例子, 打开数据文件“data5-1.sav”, 具体设置如下。

菜单选择: “分析→非参数检验→单样本(O)”, 打开“单样本非参数检验”对话框, 根据要进行的具体分析进行相应的设置。

(1) “目标”选项卡: 用于设置非参数检验的目标, 每个不同的选项对应于“设置”选项卡上不同的默认配置, 如图 5-1 所示。

① “自动比较实测数据和假设数据”: 选中此项, “设置”选项卡上的“选择检验”项将自动设置为“根据数据自动选择检验”, 系统将根据“字段”选项卡中的相应字段自动选择二项检验、卡方检验或 K-S 检验进行比较观察数据及假设检验; 如果待检验变量是具有两个不同取值的分类变量, 将用二项分布检验; 对所有其他分类字段用卡方检验; 对连续字段用 K-S 检验。

② “检验序列的随机性”: 选中此项, “设置”选项卡上的“选择检验”项将自动设置为“检验随机序列(游程检验)”。



图 5-1 “单样本非参数检验”对话框“目标”选项卡

③ “定制分析”：选中此项，“设置”选项卡上的“选择检验”项将允许手动选择要执行的检验及对选项进行设置。

(2) “字段”选项卡：用于设定待检验变量。

单击“字段”选项卡，按图 5-2 所示进行设置。该选项卡主要由以下几部分组成。



图 5-2 “单样本非参数检验”对话框“字段”选项卡

① 使用预定义角色：不须对检验字段进行设置，系统自动将数据文件“变量视图”窗口中定义为“输入”、“目标”、“两者都”角色的变量加入“检验字段 (T)”列表框中，定义为其他角色的变量将不会自动进入。

② 使用自定义字段分配：手工设定检验字段。

➢ “字段 (F)”列表框：列出数据文件中的所有字段。

➤ 检验字段：该列表框中的字段作为检验字段。将要检验的字段从“字段（F）”列表框中移入。

(3) “设置”选项卡：用于设定检验方法及对应的选项，如图 5-3 所示。



图 5-3 “单样本非参数检验”对话框“设置”选项卡

① “选择检验”：设置所进行的检验及其属性。

➤ 根据数据自动选择检验：该设置对具有两个不同取值的分类变量，将用二项分布检验；对所有其他分类字段用卡方检验；对连续字段用 K-S 检验。

➤ 定制检验：允许手动设置要执行的待定检验，其中每一项的含义见后面具体分析。

② “检验选项”：设置置信区间、显著性水平及缺失值的处理。按图 5-4 所示进行设置。



图 5-4 “单样本非参数检验—设置—检验选项”对话框

➤ 按检验排除个案：只有当检验变量中含缺失值时才删除该观测量。

- 成列排除个案：凡含有缺失值的观测量全部从分析中排除。
- ③ “用户缺失值”：设置分类字段缺失值的处理方式。
- 排除：分析过程中，把有缺失值的样本去掉，不进入分析。
- 包括：缺失值将作为一个类别。

☆说明☆

以上设置为所有单样本非参数检验的共同设置。

5.2.2 卡方检验

1. 卡方检验的概念

卡方检验（Chi-Square Test）法，也称卡方拟合优度检验，它是 K.Pearson 给出的一种最常用的非参数检验方法，用于检验观测数据是否与某种概率分布的理论数值相符合，进而推断观测数据是否来自于该分布的样本。例如，根据掷骰子试验中出现的点数，检验骰子是否均匀，即各点出现的概率是否均为 1/6。卡方检验的原假设 H_0 ：总体服从某种理论分布。此外，卡方检验还可对定性行列表资料的行列变量的独立性，以及线性相关性（线性趋势）进行分析，这部分内容可参见第 3 章“交叉表分析”中的内容。

2. 统计原理

如果从一个随机变量 X 中随机抽取若干个观测样本，这些观测样本落在 X 的 k 个互不相交的子集中的观测频数服从一个多项分布，这个多项分布当 k 趋于无穷时，就近似服从 X 的总体分布。

基于上述基本思想，对变量 X 总体分布的检验就应该从对各个观测频数的分析入手。在卡方检验中，原假设给出了在假想总体中归入每一类别的对象所占的比例。也就是说，可以从原假设推出期望的频数是多少。而卡方检验则可以判断观测的频数是否充分地接近于原假设成立时可能出现的期望频数。

为检验实际分布是否与理论分布（期望分布）一致，可采用卡方统计量，典型的卡方统计量是 Pearson 卡方统计量，其公式：

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \tag{5.1}$$

式中， k 为子集个数； n_i 为第 i 个子集的频数； n 为样本总量； p_i 为第 i 个子集的理论频率。

如果卡方值较大，则说明期望频数与观测频数分布差距较大，没有证据支持原假设；若卡方值较小，则说明期望频数与观测频数比较接近，不能拒绝原假设。

3. 分析步骤

卡方检验也属于假设检验问题，具体步骤如下。

第 1 步 提出零假设。

卡方检验的零假设 H_0 ：总体服从某种理论分布，其对立假设 H_1 ：总体不服从某种理论分布。

第 2 步 选择检验统计量。

卡方分布选择的是 Pearson 卡方统计量。已证明，当 n 充分大时，它近似地服从自由度为 $k-1$ 的卡方分布 $\chi^2(k-1)$ 。

第 3 步 计算检验统计量的观测值和概率 P 值。

SPSS 会根据式 (5.1) 自动计算 χ^2 统计值, 并依据 χ^2 分布表给出相应的相伴概率值 P 。从式 (5.1) 可知, 如果 χ^2 值较大, 则说明观测频数分布与期望频数分布差距较大; 反之, 如果 χ^2 值较小, 则说明观测频数分布与期望频数分布接近。

第 4 步 给出显著性水平 α , 作出决策。

如果显著性概率 P 值小于显著性水平 α , 则拒绝零假设, 认为样本来自的总体服从理论分布; 反之, 认为样本来自的总体分布与期望分布存在显著性差异。

☆说明☆

卡方检验过程要求检验变量的数据类型是度量标准为“名义”或“有序”的分类变量。

4. 卡方检验 SPSS 实例分析

【例 5-1】某公司质检负责人欲了解企业一年内出现的次品数是否均匀分布在一周的五个工作日中, 随机抽取了 90 件次品的原始记录, 其结果如表 5.1 所示, 问该企业一周内出现的次品数是否均匀分布在一周的五个工作日中? ($\alpha=0.05$) (数据来源: 周惠彬,《应用统计学》, 西南财经大学出版社; 参见数据文件: data5-1.sav。)

第 1 步 分析。

由于考虑的是次品是否服从均匀分布的问题, 故用卡方检验。

第 2 步 数据组织。

首先建立 SPSS 数据文件, 建立两个变量: “工作日”、“次品数”, 录入相应数据并保存。(注意: “工作日”字段是度量标准为“有序”或“名义”的字符或数值类型, “次品数”字段是度量标准为“标度”的数值类型。)

第 3 步 “次品数”字段加权处理。

通过分析“工作日”及“次品数”两个字段的含义及度量标准, 确定“工作日”为被分析字段, 而“次品数”表示各工作日出现的频数, 所以应该对“次品数”进行加权处理。执行“数据”→“个案加权”, 打开“个案加权”对话框, 按图 5-5 所示进行设置。

表 5.1 某企业产品次品抽样数据

工作日	1	2	3	4	5
次品数	25	5	8	6	6



图 5-5 “个案加权”对话框

第 4 步 单样本的非参数检验设置。

选择菜单“分析→非参数检验→单样本”, 打开如图 5-3 所示的“单样本非参数检验”对话框, 按下面的步骤进行设置。

(1) 在“目标”选项卡选择“自定义分析”。

(2) 在“字段”选项卡选择“使用自定义字段分配”，并将“工作日”字段选入“检验字段”。

(3) “设置”选项卡中选择“定制检验”，并选中“比较实测概率和假设概率（卡方检验）”，“检验选项”及“用户缺失值”保持默认选项。

第5步 卡方检验的选项设置。

在图 5-3 上单击“卡方检验”对应的“选项”按钮，打开“卡方检验选项”对话框，按图 5-6 所示进行设置。

- ① 所有类别的概率相等 (V)：表示每个分类的期望值都相同，检验变量值服从均匀分布。
- ② 定制期望概率 (C)：每个分类的期望值是不同的，需要用户自己将计算好的期望值输入“期望概率”中。

第6步 运行结果及分析。

完成以上操作步骤后，单击图 5-4 中的“运行”按钮，运行结果如图 5-7 所示，具体意义分析如下。



图 5-6 “卡方检验选项”对话框

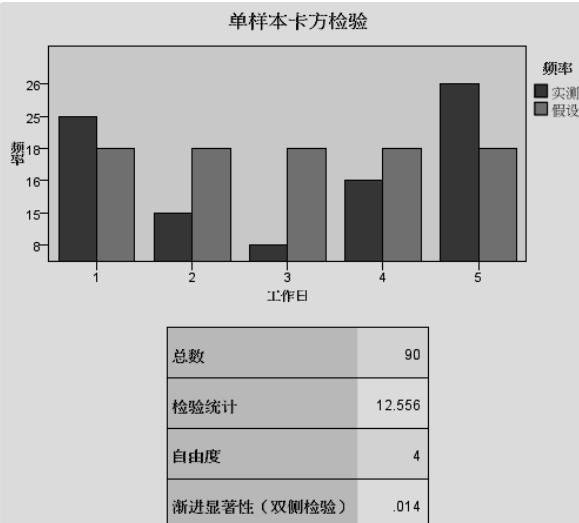
假设检验摘要			
原假设	检验	显著性	决策
1 工作日的类别以同等概率出现。	单样本卡方检验	.014	拒绝原假设。

显示了渐进显著性。显著性水平为 .05。

图 5-7 卡方检验的假设检验数据摘要

图 5-7 为单样本卡方检验的假设检验结果，根据前面的设置，给出了卡方检验的原假设为“工作日的类别以同等概率出现”，其显著性概率 $P = 0.014 < 0.05$ ，说明应拒绝原假设，因此图 5-7 的“决策”给出“拒绝原假设”的决策，认为工作日的类别是以不同概率发生的，即认为该企业一周内出现的次品数不是均匀分布在一周的五个工作日中。

双击输出文件中如图 5-7 所示的假设检验汇总表，打开如图 5-8 所示的图形，从图 5-8 可以更直观地看出工作日的次品数不是均匀分布在每一天的，其中星期一及星期五的次品数是最高的，工作效率最低；且从中可以看出共有 90 个样本，检验统计量为 12.556，渐进显著性（即 P 值）为 0.014。



1. 共有 0 个期望值小于 5 的单元格 (0%)。最小期望值为 18。

图 5-8 单样本卡方检验的模型浏览器

☆说明☆

- ◆ 如果实际情况不是均匀分布，则必须将分布数列输入“期望概率”，在列表中自定义期望概率值时，所有的期望概率值相加应该为 1。也可以输入整数值，系统会自动计算输入的每个整数值在所有输入的整数值中所占的比例，并将这个比例值视为所对应每一类的期望概率值（即系统会自动做规一化处理）。

【例 5-2】 长期以来，一批教员对某一特殊课程所给出的平均成绩等级 A : B : C : D : E 的比例都是 12 : 18 : 40 : 18 : 12，现有一名新教员，两个学期以来对这门课程给出了 22 个 A，34 个 B，66 个 C，16 个 D 和 12 个 E，如表 5.2 所示，试在显著性水平 0.05 下，确定新教员的评分形式是否与其他人一样。（数据来源：M.R.斯皮格尔，著，杨纪龙，等译，《统计学》，科学出版社；参见数据文件：data5-2.sav。）

第 1 步 分析。

由于要判断新教员所打的成绩等级分布是否服从其他教员的等级分布形式，因此采用卡方检验。

表 5.2 某新教员所打成绩等级的分布

成绩等级	A	B	C	D	E
人数	22	34	66	16	12

第 2 步 数据的组织。

数据分成两列，一列是成绩等级，其变量名为“grade”；另一列是人数，变量名为“persons”，输入数据并保存。

第 3 步 加权设置。

将变量“persons”定义为加权变量，设置方法与例 5-1 类似。

第 4 步 单样本的非参数检验设置。

设置方法与例 5-1 类似，入选“检验字段”的是“成绩等级”。

第 5 步 卡方检验的选项设置。

方法与例 5-1 相同，但在图 5-6 的“选择检验选项”中选择“定制期望概率”，按图 5-9 所示进行设置。

第 6 步 主要结果及分析。

完成以上操作步骤后，单击图 5-9 中的“确定”按钮，运行结果如图 5-10 所示，具体意义分析如下。

图 5-10 为单样本卡方检验的假设检验结果，根据前面的设置，给出了卡方检验的原假设“成绩等级的类别以指定概率发生”，即认为新教员给出的成绩等级分布与其他教员相同，其相伴概率值显著性概率 $P = 0.044 < 0.05$ ，说明应拒绝原假设，因此图 5-10 的“决策”给出“拒绝原假设”的决策。



图 5-9 “卡方检验选项”对话框

假设检验摘要			
原假设	检验	显著性	决策
1 成绩等级 的类别以指定概率出现。	单样本卡方检验	.044	拒绝原假设。
显示了渐进显著性。显著性水平为 .05。			

图 5-10 单样本卡方检验的假设检验结果

前面讲到在“非参数检验”下的子菜单“旧对话框”中保留了 SPSS 18 之前的低版本的菜单供用户使用，旧对话框的设置和新对话框区别比较大，我们在下例中采用旧对话框进行卡方检验。

【例 5-3】 为了调查某校学生的英语学习态度，我们抽取部分同学做了如下问卷调查：你认为当前大学生的英语学习态度如何？

- ① 很好 ② 较好 ③ 一般 ④ 较差 ⑤ 很差

调查结果表 5.3（部分数据）直接记录了每个学生的选题编号，请从该调查结果中判断学生的学习态度有无显著差异。（数据来源：冯岩松，《SPSS 22 统计分析应用教程》，清华大学出版社；参见数据文件：data5-3.sav）

表 5.3 大学生英语学习态度调查结果

选题编号	3	1	3	4	4	2	3	2	4	4	...
------	---	---	---	---	---	---	---	---	---	---	-----

第 1 步 分析。

该题实际判断的是学生的选项是否服从均匀分布的问题，故用卡方检验。

第 2 步 数据组织。

对于表 5.3 的数据组织形式，无法使用前面的新对话框，因为新对话框要求有两个字段：频数数和检验字段，只填一个字段在运行时将会报错。对于此例，要么我们重新组织类似表 5.1 的数据，然后加权，再用新的对话框分析；要么直接对该数据用旧对话框进行分析。下面我们利用旧对话框来分析。

第 3 步 检验变量设置。

选择菜单“分析→非参数检验→旧对话框→卡方”，按图 5-11 所示进行设置。

(1) “期望范围”选项组：用于设定需检验的变量的取值范围，在此范围之外的取值将不进入分析，包含以下两个选项。

① 从数据中获取：表示检验变量的取值范围使用数据文件的最大值和最小值之间确定的范围。

② 使用指定范围：自行制定检验的取值范围，选择该项后，可在“上限”和“下限”文本框中分别输入检验范围的上限和下限。

(2) “期望值”选项组：指定已知总体的各分类构成比，包含以下两个选项。

① 所有类别相等：设定各类别构成比例相等，即检验的总体是服从均匀分布的。

② 值：用于自行定义类别构成的比例，每输入一个值后单击“添加”按钮，系统自动将其输入右边的列表框。输入数值必须大于 0，重复以上操作直到输完为止。如果在输入中出现了错误，可以选中已输入的值，单击“更改”按钮进行修正，或单击“删除”按钮将其删除，然后重新输入。输入值时要注意输入顺序一定要和变量递增的顺序一致。

第 4 步 检验精度设置。

在图 5-11 上单击“精确”按钮，打开“精确检验”对话框并按默认设置。

第 5 步 选项设置。

在图 5-11 上单击“选项”按钮，打开“卡方检验：选项”对话框，按图 5-12 进行设置。



图 5-11 “卡方检验”对话框



图 5-12 “卡方检验：选项”对话框

(1) “统计”选项组：如果勾选“描述”和“四分位数”选项，结果将输出平均值、标准差、四分位数等描述性统计变量。

(2) “缺失值”选项组：

① 按检验排除个案：只排除检验变量中含缺失值的观察单位。

② 成列排除个案：排除所有含有缺失值的变量。

第 6 步 主要结果及分析。

完成以上操作步骤后，单击图 5-11 的“确定”按钮，运行结果如表 5.4 和表 5.5 所示。

表 5.4 中第二列为学生实际选择的频率，第三列为期望频率，而“残差”列则是第二列与第三列的差值。残差为正，说明实际频率多于期望频率，为负则反之。残差的绝对值越大，表示实际频率和期望频率的差距越大。

表 5.5 为本次卡方检验的结果表。在图 5-11 “期望值”选项组中，我们选择的“所有类别相等”，即原假设为“各选项实际频率为均匀分布”。这里渐进显著性 P 值为 0.000，小于 0.05，

拒绝原假设，说明大学生英语学习态度的 5 个选项的勾选频率与期望值差异非常显著，即学生的英语学习态度有显著差异。

表 5.4 大学生英语学习态度卡方频率分析表

你认为当前的大学生英语学习态度如何？			
	实测个案数	期望个案数	残差
1	3	33.4	-30.4
2	8	33.4	-25.4
3	84	33.4	50.6
4	52	33.4	18.6
5	20	33.4	-13.4
总计	167		

表 5.5 大学生英语学习态度卡方检验结果表

检验统计	
	你认为当前的大学生英语学习态度如何？
卡方	139.377a
自由度	4
渐近显著性	.000
a. 0 个单元格 (0.0%) 的期望频率低于 5。期望的最低单元格频率为 33.4。	

5.2.3 二项分布检验

1. 基本概念

在现实生活中有很多数据是二值的，如男性和女性、生与死，患病的有和无、产品的合格与不合格等。对于这种情况，从总体中抽取的所有可能结果，要么是对立分类中的一类，要么是另一类，通常将这样的二值用 1 和 0 表示。如果进行 n 次相同的试验，则出现两类（1 或 0）的次数可以用离散型随机变量 X 来描述。如果随机变量 X 值为 1 的概率为 P ，则 X 为 0 的概率 q 等于 $1 - P$ ，这样的分布为二项分布。

二项分布检验正是通过样本数据检验样本来自的总体是否服从指定概率为 p 的二项分布，其原假设 H_0 ：样本来自的总体与指定的二项分布无显著性差异。

2. 统计原理

二项分布检验在样本 ≤ 30 时，按式（5.2）计算概率值：

$$P\{X \leq x\} = \sum_{i=1}^x C_n^i p^i q^{n-i}$$

(5.2)

表示 n 次试验中某类出现的次数 $\leq x$ 的概率。在大样本情况下，计算的是 Z 统计量，认为在原假设下， Z 统计量服从正态分布，其计算公式：

$$Z = \frac{x \pm 0.5 - np}{\sqrt{np(1 - p)}}$$

(5.3)

式中，当 x 小于 $n/2$ 时，取加号，反之取减号； p 为检验概率； n 为样本总数。

3. 分析步骤

二项分布检验亦是假设检验问题，检验步骤同假设检验步骤。SPSS 会自动计算上述精确概率和近似概率值。如果概率值小于显著性水平 α ，则拒绝原假设，认为样本来自的总体与指定的二项分布有显著差异，反之，样本来自的总体与指定的二项分布无显著差异。

4. 二项分布检验 SPSS 实例分析

【例 5-4】有 20 名学生经过新型教学法后测试成绩如表 5.6 所示，以 90 分及以上为优秀，请检验这 20 名学生的优秀率是否达到了 10%。（参见数据文件：data5-4.sav。）

表 5.6 20 名学生的测试成绩

成绩	78	75	84	76	89	93	94	88	95	87	88	73	84	82	80	84	87	91	95	83
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

第 1 步 分析。

由于成绩仅分为优秀与非优秀两种状态，而且测试的是优秀率是否达到了 10%，故应用二项分布检验。

第 2 步 数据的组织。

数据组织成一列，其变量名为“成绩”，输入数据并保存。

第 3 步 单因素的非参数检验设置。

选择菜单“分析→非参数检验→单样本”，打开如图 5-3 所示的“单样本非参数检验”对话框，按以下步骤进行设置：

- （1）“目标”选项卡中选择“定制分析”；
- （2）“字段”选项卡中选择“使用自定义字段分配”，并将“成绩”字段选入“检验字段”；
- （3）“设置”选项卡中选择“定制检验”，并选中“比较实测二元概率和假设二元概率（二项检验）(O)”，“检验选项”及“用户缺失值”保持默认选项。

第 4 步 进行二项分布检验选项设置。

单击“二项检验”对应的“选项”按钮，打开“二项选项”对话框，按图 5-13 所示进行设置。

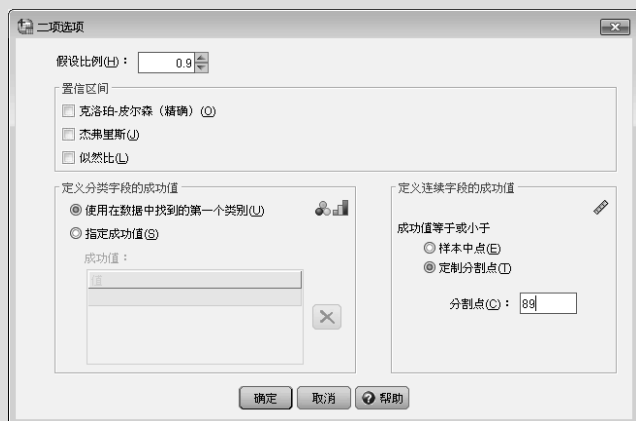


图 5-13 “二项选项”对话框

该选项主要由以下几部分组成。

- （1）“假设比例”：指定检验的原假设。输入一个在范围 0.001~0.999 之间的数值，作为待检测的第一组的概率。

- (2) 定义分类字段的成功值：当检验变量是二元变量时使用。
- ① 使用数据中找到的第一个类别：将检验字段中出现的第一个分类作为“成功”组。
 - ② 指定成功值：指定检验字段中的某一个特定分类为“成功”组。
- (3) 定义连续字段的成功值：当检验变量不是二元变量而是连续变量时使用。

- ① 样本中点：取样本的中点作为分组标准，小于或等于中点的样本为“成功”组，大于中点的样本为“失败”组。
- ② 定制分割点：在其后的“分割点”文本框内输入割点值，系统将自动将变量值小于或等于割点值的样本分为“成功”组，其余的分为“失败”组。

第 5 步 主要结果及分析。

按以上步骤设置后，运行二项分布检验得到的检验结果如图 5-14 所示。

图 5-14 为单样本二项检验结果，检验结果给出了二项检验的原假设“成绩≤89 的学生和成绩>89 的学生的比例为 0.9：0.1”，其显著性水平默认设置为 5%，显著性概率 $P=0.043<0.05$ ，因此“决策者”给出“拒绝原假设”的决策，认为“成绩 ≤89 的学生与成绩>89 的学生比例显著不等于 0.9：0.1”。

双击输出文件中如图 5-14 的假设检验汇总表，打开如图 5-15 所示图形，从图 5-15 可以更直观地看出：“成功”组堆积在“失败”组的顶部；相对于假设的分布，实际观察的分布中，成绩>89 的学生比例更高一些；且数据文件中共有 20 个样本，其检验统计量的值为 15.000，标准误差为 1.342，标准化检验统计量为-1.863，渐进显著性及精确显著性分别为 0.031 及 0.043。

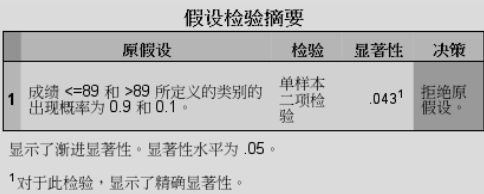
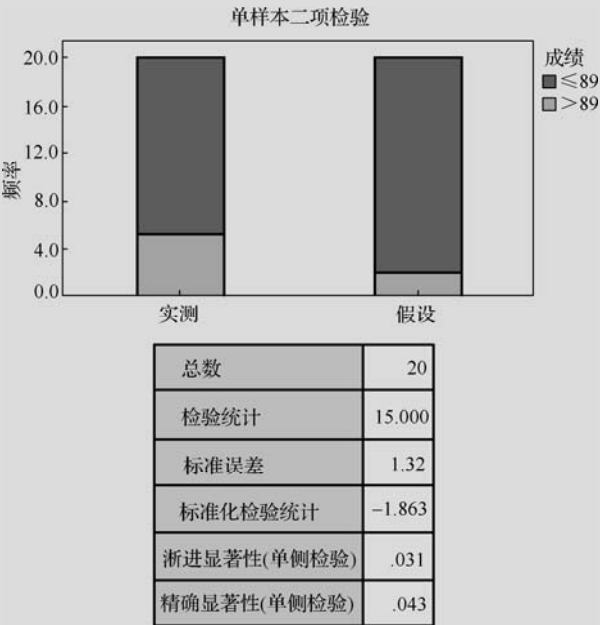


图 5-14 二项假设检验数据摘要



1. 备用假设如下：成功组中的记录比例小于假设的成功概率。

图 5-15 二项检验结果的堆积条形图

☆说明☆

- (1) 二项检验中把检验字段第1组的类别称为成功, 把检验字段剩下组的类别称为“失败”, 所以二项检验的选项中设置检验字段的成功值即设置检验字段的第一个分类组。
- (2) 这里“优秀”是指变量“成绩” ≥ 90 , 因此定义断点时不能取值90, 而应取89。在“分割点”文本框中输入的数值是“成功”组的概率, 而这里的“成功”组小于90, 因此应输入0.9, 而不是0.1。
- (3) 此题还有另一种求解方法, 就是先对变量的原始数据执行“转换→重新编码为其他变量”命令, 重新赋值产生一个二元变量, 然后对这个二元变量进行检验, 请读者自己完成。

5.2.4 游程检验

1. 基本概念

一个游程(Run)就是某序列中位于一种符号之前或之后的另一种符号持续的最大主序列, 或者说, 一个游程是指某序列中同类元素的一个持续的最大主集。游程检验(Runs Test)又称变量的随机性检验, 主要用于检验一个变量两个值的分布是否呈随机分布, 即检验前一个个案是否影响下一个个案的值, 如果没有影响, 这一组个案便是随机的。对于连续型变量的随机性检验也可转化为只有两个取值的分类变量的随机性检验。其原假设 H_0 : 变量值的分布是随机的。

例如, 30次掷硬币出现正反面的序列为000011100000110000011111100000, 如果称连在一起的0或连在一起的1为一个游程, 则共有4个0游程和3个1游程, 共7个游程($R=7$)。可以直观地理解为, 如果硬币的正反面出现是随机的, 那么在该数据序列中, 许多个1或许多个0出现的可能性将不太大, 同时, 1和0频繁交叉出现的可能性也会很小。因此, 游程数太大或太小都将表明变量值存在不随机的现象。

利用游程检验可以对次序统计量进行随机性检验, 还可以对不同的两个总体进行显著性检验。

2. 统计原理

SPSS 单样本变量随机性检验中, 利用游程数构造检验统计量。如果设 n_1 为出现1的个数, n_2 为出现0的个数, 当 n_1, n_2 较大时, 游程抽样分布的均值为 $\mu_r = \frac{2n_1n_2}{n_1+n_2}$, 方差为 $\sigma_r^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$ 。

在大样本条件下, 游程近似服从正态分布, 即

$$Z = \frac{r - \mu_r}{\sigma_r} \quad (5.4)$$

式中, r 为游程数。

3. 分析步骤

游程检验也是假设检验问题, 检验步骤同前。SPSS 会根据式(5.4)自动计算 Z 统计量, 并依据正态分布表给出对应的概率 P 值。如果概率值小于显著性水平 α , 则拒绝原假设, 认为变量的分布不是随机的, 反之认为变量值的出现是随机的。

4. 游程检验 SPSS 实例分析

【例 5-5】某股票连续 20 天的收盘价如表 5.7 所示, 在显著性水平 0.05 下, 判断此价格是否是随机的? (数据来源: M.R.斯皮格尔, 《统计学(第3版)》, 科学出版社; 参见数据文件: data5-5.sav。)

表 5.7 某股票连续 20 天的收盘价

10.375	11.125	10.875	10.625	11.500	11.625	11.250	11.375	10.750	11.000
10.875	10.750	11.500	11.250	12.125	11.875	11.375	11.875	11.125	11.750

第 1 步 分析。

由于判断的是价格是否随机分布，可用游程检验对统计量进行随机性检验。该检验的原假设 H_0 ：样本是随机的。

第 2 步 数据组织。

将这些数据组织成一列，变量名为“price”，输入数据并保存。

第 3 步 单因素的非参数检验设置。

选择菜单“分析→非参数检验→单样本”，打开如图 5-3 所示的“单样本非参数检验”对话框，按以下步骤进行设置：

(1) 在“目标”选项卡选择“定制分析”。

(2) 在“字段”选项卡中选择“使用自定义字段分配”，并将“price”字段选入“检验字段”或使用默认设置。

(3) 在“设置”选项卡中选择“定制检验”，并选中“检验序列的随机性（游程检验）”，“检验选项”及“用户缺失值”保持默认选项。

第 4 步 游程检验的选项设置。

在图 5-3 所示的“单样本非参数检验”对话框中单击“检验序列的随机性（游程检验）”对应的“选项”按钮，打开“游程检验选项”对话框，按图 5-16 所示保持默认设置。

该选项卡主要由以下几部分组成。

(1) 定义分类字段的组：当检验字段为分类字段时使用；

① 样本中只有 2 个类别：当检验字段为分类字段，且只有 2 个类别时使用；如投硬币时用 0 和 1 分别表示正、反面。

② 将数据重新编码为 2 个类别：当检验字段为分类字段，且有 2 个以上类别时使用，将检验字段中的类别重新定义为 2 个类别。在“定义第一个类别”中输入要定义在新的第一个类别中的多个原有类别值，类别之间用逗号隔开。

如用 1、2、3、4、5 分别表示某件次品是在工作日的哪一天生产的，这是一个分类字段，且该字段超过了 2 个类别。如果将 5 个类别重新编码，原有的 1、2 为一个类别，3、4、5 为另一个类别，便将原有的 5 个类别重新定义为了 2 个类别。

(2) 定义连续字段的分割点：当检验字段为连续字段时设置割点。

① 样本中位数：以样本的中位数为割点。

② 样本平均值：以样本均值为割点。

③ 定制：用户自定义割点值。

④ 分割点：选中“定制”后才可使用，在对应的文本框中输入自定义的割点值。



图 5-16 “游程检验选项”对话框

本例中数据为连续数据，以样本中位数为割点，将各样本值重新编码为 0 或 1 的 2 个类别的数据序列。

第 5 步 主要结果及分析。

完成以上操作步骤后，单击图 5-3 中的“运行”按钮，运行结果如图 5-17 所示，具体意义分析如下。

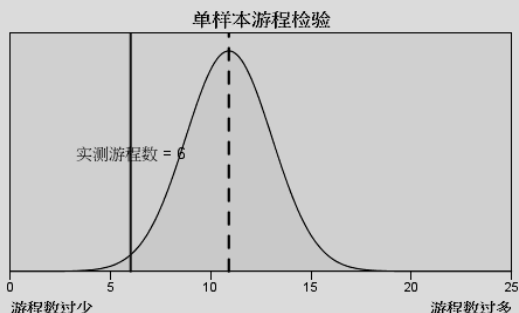
假设检验摘要			
	原假设	检验	显著性
1	股价 ≤ 11.250 和 >11.250 所定义的 值序列是随机序列。	单样本 游程检 验	.041
决策			
拒绝原 假设。			

显示了渐进显著性。显著性水平为 .05。

图 5-17 游程检验的数据摘要

图 5-17 为单样本游程检验的假设检验数据摘要，给出了本例游程检验的原假设“由股价 ≤ 11.25 和 >11.25 定义的值的序列是随机序列”，其显著性水平为 0.05，显著性概率 $P=0.041<0.05$ ，因此“决策”给出“拒绝原假设”的决策，认为“由股价 ≤ 11.25 和 >11.25 定义的值”的序列不是随机序列。

双击输出文件中如图 5-17 所示的“假设检验摘要”表，打开如图 5-18 所示图形，显示了游程检验的图表和检验表，其中检验统计量为 6.000，标准误差为 2.153，显著性概率 P 值为 0.041，图表显示了以垂直线标记的观察到的游程数的正态分布。



总数	20
检验统计	6.000
标准误差	2.153
标准化检验统计	-2.043
渐进显著性（双侧检验）	.041

图 5-18 游程检验结果

☆说明☆

- (1) 这种题检验的过程是先以中位数为标准，将这 20 个数分成 0 和 1 组成的序列，小于中位数的为 0，大于和等于中位数的为 1，则该序列为 0000111100001111101，SPSS 是根据这个序列求出其中的游程再做检验的；
- (2) 上例的分割点选择的是中位数，亦可选择均值或众数，请读者自己完成。

【例 5-6】 用甲、乙两台机床生产同一型号的滚珠，今从甲、乙两台机床生产的滚珠中分别抽取 8 个和 7 个，测得的直径数据如表 5.8 所示，问这两台机床生产的滚珠在显著性水平 0.05 下是否有差别？（数据来源：耿修林，《应用统计学》，科学出版社；参见数据文件：data5-6.sav。）

表 5.8 两台机床生产滚珠的直径数据表

机床甲	9.9	9.6	10	9.6	9.8	10.1	9.5	9.7
机床乙	9.3	9.4	10.2	10.2	10.1	9.7	9.4	

第 1 步 分析。

由于不知道两台机床生产的滚珠的分布，本问题相当于检验原假设 H_0 ：两总体（机床甲和机床乙）有相同的分布，用游程检验可以处理这个问题，对不同的两个总体进行显著性检验。

第 2 步 数据组织。

分成两列数据，其一是所有滚珠的直径，变量名为“diameter”，度量标准为“度量”；其二是机床，变量名为“machine”（变量值 1 表示机床甲，2 表示机床乙），度量标准为“序号”，输入数据并保存。

第 3 步 数据排序。

执行“数据→个案排序”命令，打开“个案排序”对话框，选择变量“diameter”移入“排序依据”框中，按照“升序”排序。

第 4 步 单因素的非参数检验设置。

选择菜单“分析→非参数检验→单样本”，打开如图 5-3 所示的“单样本非参数检验”对话框，按以下步骤进行设置：

- （1）在“目标”选项卡选择“定制分析”；
- （2）在“字段”选项卡中选择“使用自定义字段分配”，并将“机床”字段选入“检验字段”；
- （3）在“设置”选项卡中选择“定制检验”，并选中“检验序列的随机性（游程检验）”，“检验选项”及“用户缺失值”保持默认选项。

第 5 步 游程检验的选项设置。

在图 5-3 所示界面上单击“检验序列的随机性(游程检验)”对应的“选项”按钮，打开“游程检验选项”对话框，因为检验字段“机床”的度量标准为“序号”，且“机床”字段中仅有 2 个类别“1”和“2”，故选择“定义分类字段的组”选项组中的“样本中只有 2 个类别”选项，设置界面同图 5-16。

第 6 步 主要结果及分析。

完成以上操作步骤后，单击图 5-3 中的“运行”按钮，运行结果如图 5-19 所示，具体意义分析如下。

图 5-19 为单样本游程检验的假设检验数据摘要，给出了本例游程检验的原假设“由机床= (2.00) 和 (1.00) 定义的值序列是随机序列”，即原假设认为两机床生产的滚珠无显著性差异。其显著性水平为 0.05，显著性概率 $P=0.110>0.05$ ，因此，“决策”给出“保留原假设”的决策，认为两机床生产的滚珠无显著性差异。

假设检验摘要			
	原假设	检验	显著性 决策
1	机床 = (2) 和 (1) 所定义的值序列是随机序列。	单样本游程检验	.110 保留原假设。

显示了渐进显著性。显著性水平为 .05。

图 5-19 游程检验的假设检验数据摘要

双击输出文件中如图 5-19 所示的“假设检验摘要”表，打开如图 5-20 所示的游程检验结果图表和检验表，具体分析同例 5-5。

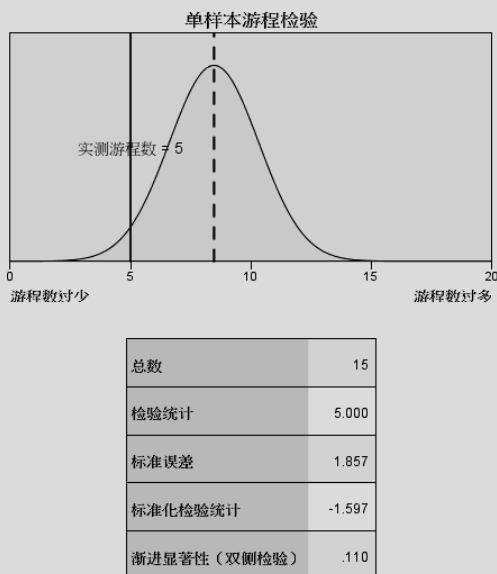


图 5-20 游程检验结果

☆说明☆

- ◆ 本题是在“单样本”菜单中检验两样本总体的差异：假设容量为 m 和 n 的两组样本，其元素分别为 a_1, a_2, \dots, a_m 及 b_1, b_2, \dots, b_n ，为了判断两组样本是否取自同一总体，首先合并两组样本，并排序，使其成为一递增序列，序列长度为 $m + n$ 。再检验序列是否是随机的，若序列是随机的，可知样本间不存在显著差异，即来自同一个总体，否则来自不同的总体。

5.2.5 单样本 K-S 检验

1. 基本概念

K-S 检验利用样本数据推断样本来自的总体是否服从某一指定分布，是一种拟合优度的检验方法，适用于探索连续性随机变量的分布。

单样本 K-S 检验可以将一个变量的实际频数分布与正态分布、均匀分布、泊松分布和指数分布进行比较。

2. 统计原理

单样本 K-S 检验的原假设 H_0 ：样本来自的总体与指定的理论分布无显著性差异。其检验的基本思路是：首先，在原假设成立的前提下，查分布表得到相应的理论累计概率分布函数 $F(x)$ ；其次，利用样本数据计算各样本数据点的累计概率，得到检验累计概率分布函数 $S(x)$ ；再次，计算 $F(x)$ 和 $S(x)$ 的差值序列 $D(x)$ ；最后，计算差值序列中的最大绝对差值，即 $D = \max(|S(x_i) - F(x_i)|)$ 。显而易见，如果样本总体的分布与理论分布差异不明显，则 D 不应该较大；否则样本总体的分布与理论分布差异就较大。

3. 分析步骤

单样本的 K-S 检验也是假设检验问题，SPSS 会自动计算 $\sqrt{n}D$ 统计量（这是大样本下的统计量）的显著性概率 P 值。如果 P 小于显著性水平 α ，则应拒绝原假设，认为样本来自的总体与指定的分布有显著差异；如果 P 大于显著性水平 α ，则应接受原假设，认为样本来自的总体与指定的分布无显著性差异。

4. 单样本 K-S 检验 SPSS 实例分析

【例 5-7】 在一批相同型号的电子元件中随机取 10 个做寿命试验，测得它们的使用寿命（单位：h）如下：420、500、920、1380、1510、1650、1760、2100、2320、2350。检验在显著性水平 $\alpha = 0.05$ 下，该批电子元件的寿命是否服从指数分布。（参见数据文件：data5-7.sav。）

第 1 步 分析。

由于是检验样本来自的总体是否服从指数分布的问题，故应该用单样本的 K-S 检验。

第 2 步 数据组织。

将以上寿命数据设置为一列，变量名为“life”，标签为“寿命”，将以上数据输入并保存。

第 3 步 单因素的非参数检验设置。

选择菜单“分析→非参数检验→单样本”，打开如图 5-3 所示的“单样本非参数检验”对话框，按以下步骤进行设置：

（1）在“目标”选项卡中选择“定制分析”；

（2）在“字段”选项卡中选择“使用自定义字段分配”，并将“寿命”字段选入“检验字段”，系统将对“寿命”字段进行相应检验；

（3）在“设置”选项卡中选择“定制检验”，并选中“检验实测分布和假设分布（柯尔莫戈洛夫-斯米诺夫检验）”，“检验选项”及“用户缺失值”保持默认选项。

第 4 步 K-S 检验的选项设置。

在图 5-3 所示界面上单击“检验实测分布和假设分布（柯尔莫戈洛夫-斯米诺夫检验）”对应的“选项”按钮，打开“柯尔莫戈洛夫-斯米诺夫检验选项”对话框，按图 5-21 所示进行设置。



图 5-21 “柯尔莫戈洛夫-斯米诺夫检验选项”对话框

从图 5-20 可以看出，K-S 检验可以检验正态分布、均匀分布、泊松分布和指数分布四种，在每种分布的设置中，需要设置其分布的参数或平均值，可以使用样本数据，也可以自己定制。

本例中检验的是“样本来自的总体是否服从指数分布”问题，所以选择“指数”假设分布，并且选择“样本平均值”作为指数分布中的平均值。

第 5 步 运行结果及分析。

完成以上操作步骤后，单击图 5-3 中的“运行”按钮，运行结果如图 5-22 所示，具体意义分析如下。

假设检验摘要				
	原假设	检验	显著性	决策
1	寿命 的分布为指数分布，平均值为 1,491。	单样本柯尔莫戈洛夫-斯米诺夫检验	.315	保留原假设。

显示了渐进显著性。显著性水平为 .05。

图 5-22 单样本非参数的 K-S 检验数据摘要

图 5-22 为单样本非参数的 K-S 检验数据摘要，给出了本例 K-S 检验的原假设“寿命的分布为指数分布，平均值为 1491”，其显著性水平为 0.05，显著性概率 $P = 0.315 > 0.05$ ，因此“决策”给出“保留原假设”的决策，即认为寿命的分布服从指数分布。

双击输出文件中如图 5-22 的“假设检验汇总”表，打开如图 5-23 所示 K-S 检验结果的图表和检验表，从图表和检验表中可以更详细、更直观地理解决策者给出的决策。

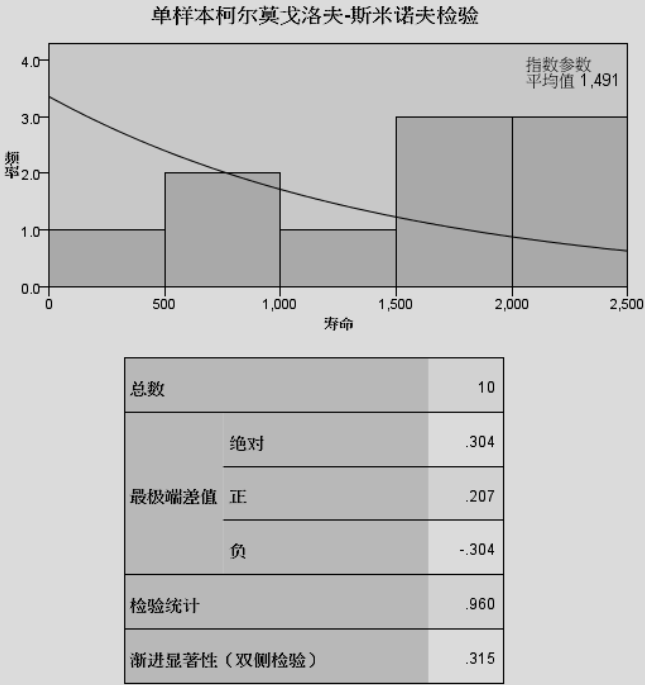


图 5-23 单样本非参数 K-S 检验结果

图 5-23 中，显示“总数”为 10 个同类型的电子元件；“最极端差值”表示样本数据与理论数据的最大差值，最大绝对值之差为 0.304，最大正差值为 0.207，最大负差值为 -0.304，最大正

负差值的大小可以判别理论分布与经验分布的差距,为计算检验统计量提供了直观的分析;“检验统计”为 K-S 正态统计量的值,这里为 0.96;“渐进显著性(双侧检验)”的值为 0.315,大于给定的显著性水平 0.05,所以应接受原假设,认为这 10 个电子元件的使用寿命分布与指数分布无显著性差异,也就是说这 10 个电子元件的寿命服从指数分布。

5.3 独立样本非参数检验

5.3.1 基本概念及统计原理

1. 基本概念

独立样本的非参数检验是通过将两组或多组独立样本的分析,推断来自两个或多个总体的分布是否存在显著性差异。之所以称为非参数检验,是因为检验过程不需要已知总体的分布,也不需要已知总体的参数。

2. 统计原理

SPSS 提供了多种独立样本的非参数检验方法,主要包括曼-惠特尼 U 检验、柯尔莫戈洛夫-斯密诺夫(2 个样本)检验、检验序列的随机性(针对 2 个样本的瓦尔德-沃尔福威茨检验)检验、摩西极端反应(2 个样本)检验、克鲁斯卡尔-沃利斯单因素 ANOVA 检验(k 个样本)、中位数检验(k 个样本)、有序备用项检验(针对 k 个样本的约克海尔-塔帕斯特拉检验),前 4 种检验是针对两独立样本的非参数检验,后 3 种检验是针对 k 个独立样本的非参数检验。

(1) 曼-惠特尼 U 检验

曼-惠特尼 U 检验即 Mann-Whitney U 检验,该检验也称威尔科克森(Wilcoxon W)等级之和检验,该检验是一种检验平均秩的差检验,可用来检验两个独立样本是否来自同一总体,它是最强的非参数检验之一,用该法进行检验时,首先将两个样本混合在一起,并对所有个案做升序排列,计算样本 1 的每个观测值大于样本 2 的每个观测值的次数,再计算样本 2 的每个观测值大于样本 1 的每个观测值的次数,分别用 U_1 和 U_2 表示,若 U_1 和 U_2 比较接近,则说明两个样本是来自相同分布的总体;若 U_1 和 U_2 差异较大,则说明两个样本来自不同的总体。检验的样本要求是连续的数据。

(2) 柯尔莫戈洛夫-斯密诺夫(2 个样本)检验

柯尔莫戈洛夫-斯密诺夫检验即 Kolmogorov-Smirnov Z 检验,简称 K-S 检验,5.2 节中我们做了单样本的 K-S 检验,在这里用来检验两组样本秩分累计频数和累计频率的差异。计算两组样本的秩分累计频数和每个点上的累计频数,将两组样本的累计频率相减,得到一组差值序列,通过检验该差值序列总和的大小来检验两个独立样本分布的差异性。检验的样本数据要求是比率数据,如果数据文件的数据不是比率数据,SPSS 过程将视其为比率数据。

(3) 检验序列的随机性(针对 2 个样本的瓦尔德-沃尔福威茨检验)检验

检验序列的随机性检验即 Wald-Wolfowitz runs 检验,简称 W-W 游程检验,是一种对两组样本秩分排列的游程检验。两个独立样本的游程检验与单样本的游程检验基本思想是一致的,不同之处在于如何得到游程数据,其方法是将两个独立样本各个案依据其分组号分别用“0”和“1”进行编号(用“0”表示第一组,“1”表示第二组),然后混合成为一个样本,并按每个个案观测值从小到大的顺序将个案重新排列,最后按每个个案分组的编号计算游程数。通过对该序列游程的检验,判断两样本是否来自同一总体。如果游程数相当大,则说明两个样本的排列是随机的,它们之间的大小是交叉出现的,即可认为两样本来自同一总体;反之,如果游程数太小,则两样本不是来自同一总体。

(4) 摩西极端反应(2个样本)检验

摩西极端反应检验即 Moses extreme reaction 检验,该检验将两个样本混合后排序,求出全部数据的秩分变量,以一个样本为控制样本,另一个样本为实验样本,以控制样本做对照,检验实验样本是否存在极端反应。首先将两组样本混合并按升序排列,然后找出控制样本最低秩和最高秩之间(即跨度)所包含的实验样本的个案数。为控制极端值对分析结果的影响,也可以先去掉样本两个最极端的观测值后再求跨度。如果跨度很小,表明两个样本无法充分混合,可以认为实验样本存在极端反应。要求检验的样本数据是连续数据。

(5) 克鲁斯卡尔-沃利斯单因素 ANOVA 检验(k 个样本)

克鲁斯卡尔-沃利斯单因素 ANOVA 检验即 Kruskal-Wallis H 检验,该检验用来检验 k 个独立样本是否来自不同总体,若这 k 个样本服从相同分布,则在样本容量不太小的情况下,式(5.5)所表示的统计量 H 服从自由度为 $k-1$ 的 χ^2 分布。

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1) \quad (5.5)$$

式中, k 为样本数; n_j 为第 j 个样本中的个案数; N 为所有样本的个案数之和; R_j 为第 j 个样本(列)中的秩和。该检验是 Mann-Whitney U 检验的推广,它不要求数据服从正态分布,因此在一定情况下可以代替 F 检验。

(6) 中位数检验(k 个样本)

中位数检验即 Median 检验,该检验用来检验 K 个独立样本是否来自同一总体,或者来自具有相同中位数的一些总体,进行检验时,根据式(5.6)计算统计量 χ^2 值:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (5.6)$$

式中, O_{ij} 为第 j 列第 i 行的个案数; E_{ij} 为假设成立时,第 j 列第 i 行的个案数。如果 k 个独立样本来自同一总体,则统计量近似服从自由度为 $k-1$ 的 χ^2 分布。当个案具有很多相同等级或数据具有二分特性时,用该检验方法较为合适。

(7) 有序备用项检验(针对 k 个样本的约克海尔-塔帕斯特拉检验)

有序备用项检验即 Jonckheere-Terpstra 检验,它与 Mann-Whitney U 检验相似,也是计算一组样本的观测值小于其他组样本观测值的个数。其统计公式:

$$J - T = \sum_{i < k} U_{ij} \quad (5.7)$$

式中, U_{ij} 为第 i 组样本观测值小于第 j 组样本观测值的个数。

3. 分析步骤

独立样本非参数检验同样属于假设检验问题,如果所计算的显著性概率 P 值小于显著性水平 α ,则应拒绝原假设(H_0 假设:几组样本所来自的独立分布总体无显著性差异),认为几组样本来自同一总体;否则应拒绝原假设,认为几组样本所来自的独立分布总体有显著性差异,即不是同一分布。

5.3.2 独立样本非参数检验 SPSS 实例分析

【例 5-8】某公司希望了解两种品牌汽油 A 和 B 每加仑的行驶里程是否有区别,表 5.9 是两种品牌汽油每加仑的行驶里程数,在显著性水平 $\alpha = 0.05$ 下,判断两个品牌间是否存在显著性差异。(数据来源: M.R.斯皮格尔,《统计学(第3版)》,科学出版社;参见数据文件: data5-8.sav.)

表 5.9 两种品牌汽油每加仑的行驶里程数

A	30.4	28.7	29.2	32.5	31.7	29.5	30.8	31.1	30.7	31.8
B	33.5	29.8	30.1	31.4	33.8	30.9	31.3	29.6	32.8	33

第 1 步 分析。

由于是两种品牌的汽油，可以认为是两组独立样本，但行驶里程数不知道服从何种分布，可用两独立样本的非参数检验进行分析。

第 2 步 数据组织。

由于独立样本的非参数检验所检验的数据只有一列，故应将 A、B 数据组织成一列，其变量名为“mileage”，定义其度量标准为“度量”，角色为“目标”；另定义一列来区分 A 和 B 品种，作分组变量，变量名为“kind”，度量标准为“序号”，角色为“输入”，设置该字段的值标签用 1 表示 A 的数据，用 2 表示 B 的数据，设置 mileage 及 kind 字段的标签分别为“行驶里程”及“汽车品种”，将数据输入并保存。

第 3 步 进行独立样本的非参数检验设置。

选择菜单“分析→非参数检验→独立样本”，打开“非参数检验：两个或两个以上的独立样本”对话框，该对话框包括 3 个选项卡，具体设置如下。

(1) “目标”选项卡：设置独立样本非参数检验的目标，每个目标对应“设置”选项卡上的一个不同默认配置。按图 5-24 所示进行设置。

- ① 在各个组之间自动比较分析：根据文件中数据选择二样本或 K 样本对应的检验自动比较不同组间的分布；选择此项，“设置”选项卡上默认选择“根据数据自动选择检验”选项。
- ② 在各个组之间比较中位数：自动用 K 样本中位数检验比较不同组间的中位数；选择此项，“设置”选项卡上默认选择“在各个组之间比较中位数”选项。
- ③ 定制分析：允许在“设置”选项卡对执行的检验及其选项进行重新设置。



图 5-24 “独立样本非参数检验：目标”选项卡

(2) “字段”选项卡：用于设定待检验变量，具体设置如图 5-25 所示。

其中，“使用预定义角色”及“使用自定义字段分配”选项与单样本非参数检验中含义相同。因为独立样本非参数检验针对的是多组（两组或 K 组）独立样本，在数据组织时，将多组

样本组织在同一字段中，另增加一分组字段，所以，在具体设置中将增加的分组字段移入图 5-25 中的“组”文本框中。

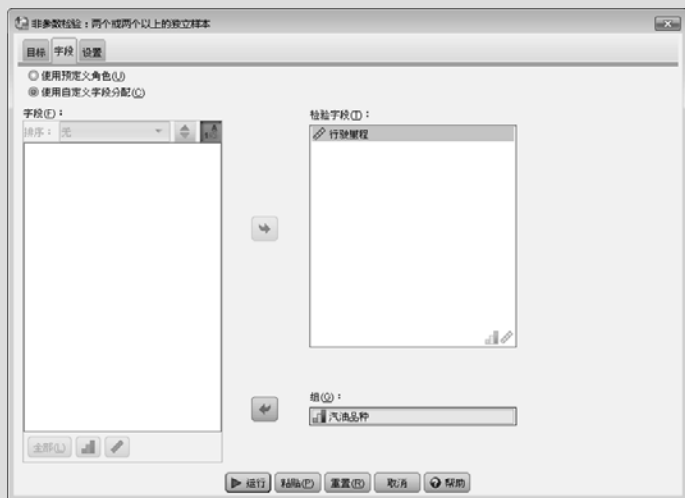


图 5-25 “独立样本非参数检验：字段”选项卡

(3) “设置”选项卡：用于设定检验方法及对应的属性，具体设置如图 5-26 所示。

① 根据数据自动选择检验：该设置对待检验变量是具有两个不同取值分类变量的情况将用两独立样本非参数检验，对具有多个不同取值的分类变量，将用 K 独立样本非参数检验，并且结合“目标”选项卡的设置，自动选择具体的检验方法。

② 定制检验：允许手动设置要执行的待定检验；本例待检验变量只有两个不同分类，故采用两独立样本非参数检验方法，将四种检验方法全部选上，表示分别用以上四种方法做检验。



图 5-26 独立样本非参数检验“设置”选项卡

第4步 主要结果及分析。

完成以上操作步骤后，单击图 5-26 中的“运行”按钮，运行结果如图 5-27 所示，具体意义分析如下。

图 5-27 为两独立样本非参数检验的数据摘要，因为在“设置”选项卡选择了四种检验方法，

所以在数据摘要中输出了四种检验方法的数据摘要。这里主要分析“曼-惠特尼 U”检验，其他检验的分析类似。

由图 5-27 可知，因为显著性水平为 0.05，而显著性概率 $P=0.165>0.05$ ，所以接受原假设，可以得知两个品牌汽油的行驶里程无显著差异。

双击输出文件中的“假设检验汇总”表，打开图 5-28 所示的检验结果，其中曼-惠特尼 U 统计量值等于 69.000；威尔科克森 W 统计量值为 124.000；标准差为 13.229；标准化检验统计量为 1.436；双侧检验的相伴概率为 0.151，大于 0.05，说明两种汽油无显著性差异；精确检验的显著性概率为 0.165，也说明两种汽油无显著性差异。

假设检验摘要				
	原假设	检验	显著性	决策
1	在 汽油品种 的类别中，行驶里程 的分布相同。	独立样本瓦尔德-沃尔福威茨游程检验	.128 ¹	保留原假设。
2	在 汽油品种 的类别中，行驶里程 的范围相同。	独立样本莫斯极端反应检验	.500 ¹	保留原假设。
3	在 汽油品种 的类别中，行驶里程 的分布相同。	独立样本曼-惠特尼 U 检验	.165 ¹	保留原假设。
4	在 汽油品种 的类别中，行驶里程 的分布相同。	独立样本柯尔莫戈洛夫-斯米诺夫检验	.400	保留原假设。

显示了渐进显著性。显著性水平为 .05。

¹对于此检验，显示了精确显著性。

图 5-27 两独立样本非参数检验数据摘要

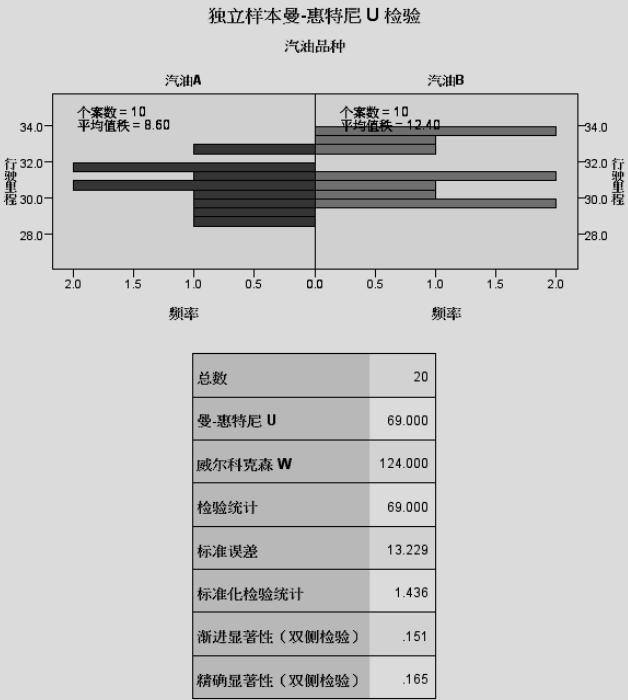


图 5-28 两独立样本非参数检验结果

☆说明☆

- ◆ 以上4种方法均为两独立样本的非参数检验方法，从例5-8分析的结果来看，检验的结果是一致的，都说明两种汽油无显著性差异。当然，对于有些数据4种方法得到的结论可能不一样，这就应根据不同的领域，选择该领域通用的处理方法来分析。

5.4 相关样本的非参数检验

5.4.1 基本概念及统计原理

1. 基本概念

相关样本的非参数检验是在对总体分布不了解的情况下，对样本所在的相关配对总体的分布是否存在显著性差异进行检验。该检验一般用于对同一研究对象（或配对对象）分别给予 K 种不同处理或处理前后的效果比较，前者推断 K 种效果有无显著性差异，后者推断某种处理是否有效。例如，5个人同时分析同一种物质中某种化学成分的含量；使用两台台秤称 N 件物品的重量；一批运动员训练前和训练后的成绩比较等。

相关样本的非参数检验对单个总体的分布不做要求，但必须是成对数据，通过比较对应样本观测值之间的差异来检验总体的差异。

在 SPSS 中，相关样本的非参数检验包括威尔科克森匹配对符号秩检验、符号检验、麦克尼马尔检验和边际齐性检验、傅莱德曼双因素按秩 ANOVA 检验、肯德尔协同系数检验和柯克兰 Q 检验，前4种检验针对两相关样本的非参数检验，后3种检验针对 K 相关样本的非参数检验。

2. 统计原理

(1) 威尔科克森匹配对符号秩检验：即 Wilcoxon 符号平均秩检验，该检验方法要求检验变量为两个连续变量，将一个样本观测值减去另一个样本相应的观测值，记下差值的符号和绝对值，将绝对值按升序排列，给出秩分，然后分别计算正值的秩分的平均值及总和、负值的秩分的平均值及总和。比较正值秩分和负值秩分的平均值与总和的差异。其计算公式：

$$Z = \frac{T - \mu_T}{\sigma_T} \quad (5.8)$$

式中， T 为检验统计量； $\mu_T = \frac{n(n+1)}{4}$ ； $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$ ； n 为样本容量。

(2) 符号检验：即 Sign 检验，该检验对不适合用定量测量而需要将每一对数据分出等级的测量最为适用，测量特征是用正、负号而不是用定量测量。符号检验是将两组样本中对应的观测值相减，分别得出正差值和负差值，比较正差值和负差值的个数的差异。这是一种对正负差值的二项分布检验。

(3) 麦克尼马尔检验：即 McNemar 检验，也称为变量显著性检验，该检验用于检验变量前后变化的显著性，要求数据是二分变量，基本方法采用二项分布检验，它通过对两组样本前后变化的频率，计算二项分布的概率值，使用 χ^2 值来判断两相关样本前后变化的差异。

(4) 边际齐性检验：即 Marginal Homogeneity 检验，也称为边缘一致性检验，该检验是对 McNemar 检验的推广。检验的两个数据变量不再为二分变量，而可以是多值的分类变量，方法是将先后测量的两样本进行 χ^2 检验。

(5) 傅莱德曼双因素按秩 ANOVA 检验：即 Friedman 双向等级方差分析，将各样本按降序从大到小并排序，得到 K 个样本的 K 列数据，然后对每行的 K 个观测值求秩分，通过各样本的总秩分和平均秩分来判断各样本的分布是否存在显著性差异。其计算公式：

$$\chi^2 = \frac{12}{bk(k+1)} \sum_{i=1}^k \left(R_i - \frac{b(k+1)}{2} \right)^2 \tag{5.9}$$

式中， b 表示样本观测值的数目； k 表示样本个数； R_i 表示第 i 组样本的秩总和。

(6) 肯德尔协同系数检验：即 Kendall's W 检验，该方法是计算协同系数 W ，以分析 K 个相关样本是否来自同一总体或具有相同的分布。 W 值介于 $0 \sim 1$ 之间， $W = 0$ 表明极度不一致； $W = 1$ 表明完全一致。Kendall's W 检验的样本必须是定序变量。

与 Friedman 双向等级方差分析一样，对每个样本的 K 个观测值求秩分，然后计算各个变量的平均秩分，并由此计算 W 值和其相应的 χ^2 值。

(7) 柯克兰 Q 检验：即 Cochran's Q 检验，也称为 Cochran 二变量检验，该检验计算 Cochran Q 系数，以分析 K 个相关样本是否来自同一总体或具有相同分布。该检验是 McNemar 检验的推广，检验的变量应该是二变量（如“是”或“否”），检验的是多个二变量取值的分布是否相同。

3. 分析步骤

相关样本非参数检验同样属于假设检验问题，如果所计算的显著性概率 P 值小于显著性水平 α ，则应拒绝原假设（ H_0 假设：几组样本所来自的独立分布总体无显著性差异），认为几组样本来自同一总体；否则应拒绝原假设，认为几组样本所来自的独立分布总体有显著性差异，即不是同一分布。

5.4.2 相关样本的非参数检验 SPSS 实例分析

【例 5-9】 某企业提出了一项新工艺，为了检验新工艺是否能降低单位成本，随机抽取 16 个工人，分别用新旧工艺生产产品，测得单位成本资料如表 5.10 所示，请在显著性水平 0.05 下检验新工艺是否降低了成本？（数据来源：周惠彬，《应用统计学》，西南财经大学出版社；参见数据文件：data5-9.sav。）

表 5.10 新旧工艺的成本情况表

new	25	12	14	22	21	17	22	16	17	18	19	24	22	15	22	23
old	18	17	16	19	24	19	28	18	22	24	22	30	25	20	24	21

第 1 步 分析。

由于是同一批工人和同一批机器，只不过是采用新旧不同的两种工艺，对某个工人来讲，其先后的成本是相关的，同时也不知数据的分布情况，故应用两相关样本的非参数检验。

第 2 步 数据组织。

数据分成两列，第一列为新工艺的成本，变量名为“new”，定义其度量标准为“度量”，角色为“两者都”；第二列为旧工艺的成本，变量名为“old”，定义其度量标准为“度量”，角色为“两者都”，将数据输入并保存。

第 3 步 进行相关样本的非参数检验设置。

选择菜单“分析→非参数检验→相关样本”，打开“非参数检验：两个或两个以上的相关样本”对话框，该对话框包括 3 个选项卡，具体设置如下。

(1) “目标”选项卡：用于设置相关样本非参数检验的目标，每个目标对应“设置”选项卡上一个不同的默认配置，本例中选择“定制分析”。

① 自动比较实测数据和假设数据：自动根据数据文件中的数据选择不同的检验方法。

② 定制分析：允许在“设置”选项卡对执行的检验及其选项进行调控。

(2) “字段”选项卡：用于设定待检验变量，其中“使用预定义角色”及“使用自定义字段分配”与单样本非参数检验中含义相同，这里可以使用默认选择“使用预定义角色”。

(3) “设置”选项卡：用于设定检验方法及对应的属性，具体设置如图 5-29 所示。

① 根据数据自动选择检验：该设置根据待检验变量的“度量标准”、待检验变量的个数并结合“目标”选项卡的设置，自动选择对应的检验方法。当待检变量为两个时，若为分类变量，用 McNemar 检验，若为连续变量，用 Wilcoxon 检验；当待检变量超过两个时，若为分类变量，用 Cochran 检验，若为连续变量，用 Kendall 协同系数及 Friedman 检验。

② 定制检验：允许手动设置要执行的检验；选中要执行的检验，运行后，在输出结果中可看到对应检验的结果，每种检验的具体用法前面已经给出，这里只给出每种检验的检验条件。

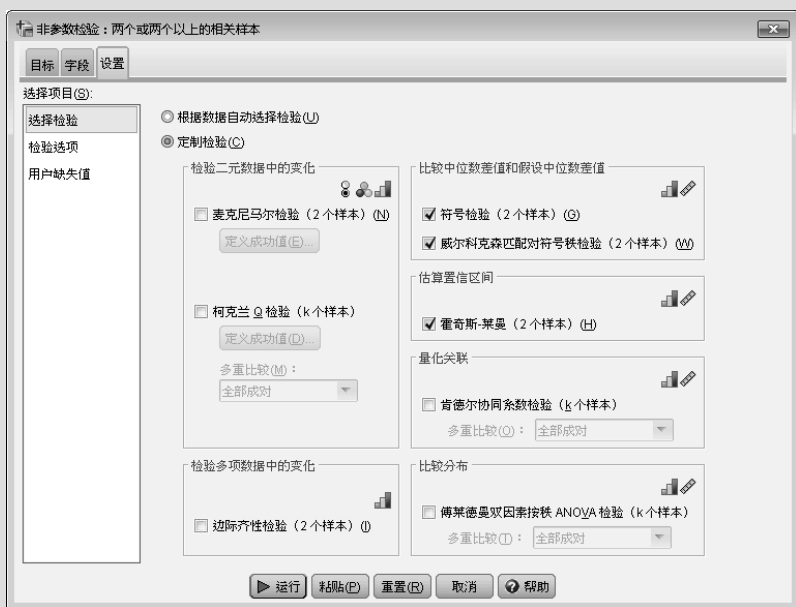


图 5-29 相关样本非参数检验“设置”选项卡

- 麦克尼马尔检验（2 个样本）(N)：要求数据为两个二分类字段。
- 柯克兰 Q 检验（k 个样本）：要求数据为两个或更多个分类字段。
- 边际齐性检验（2 个样本）：要求数据为两个有序字段。
- 符号检验（2 个样本）及威尔科克森匹配对符号秩检验（2 个样本）(W)：要求数据为两个连续变量。
- 霍奇斯-莱曼（2 个样本）：用于估计统计量的置信区间，要求待检验变量为连续变量。
- 肯德尔协同系数检验（k 个样本）：要求数据为两个或更多个连续字段。
- 傅莱德曼双因素按秩 ANOVA 检验（k 个样本）：要求数据为两个或更多个连续字段。

本例待检变量为两个连续变量，故采用两相关样本非参数检验方法，选择符号检验和威尔科克森匹配对符号秩检验。

第 4 步 主要结果及分析。

完成以上操作步骤后，单击图 5-29 中的“运行”按钮，运行结果如图 5-30 所示，具体意义分析如下。

假设检验摘要				
	原假设	检验	显著性	决策
1	新工艺 与 旧工艺 之间的差值的中位数等于 0。	相关样本符号检验	.021 ¹	拒绝原假设。
2	新工艺 与 旧工艺 之间的差值的中位数等于 0。	相关样本威尔科克森带符号秩检验	.031	拒绝原假设。

显示了渐进显著性。显著性水平为 .05。

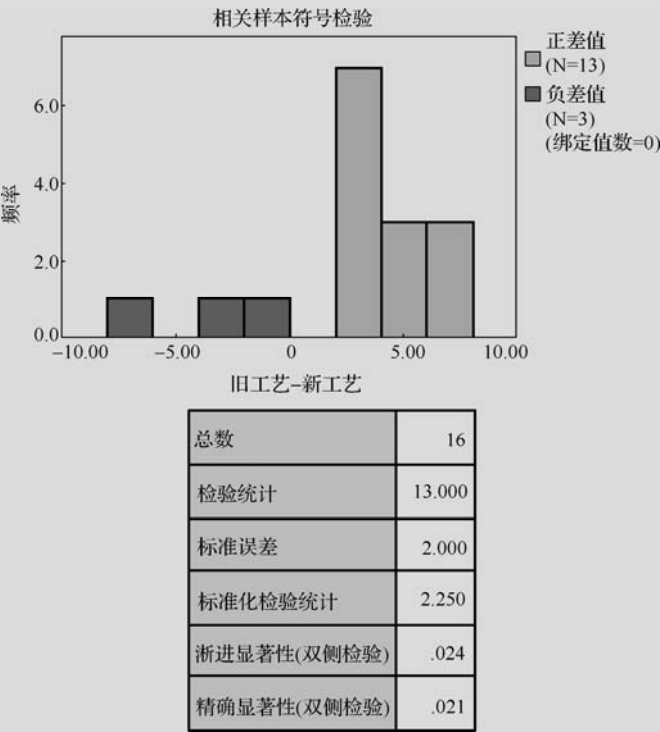
¹ 对于此检验，显示了精确显著性。

图 5-30 两相关样本非参数检验数据摘要

图 5-30 为两相关样本非参数检验的数据摘要，因为在“设置”选项框选择了“比较中位数差值和假设中位数差值”中的两种检验方法，所以在数据摘要中输出了两种检验方法的数据摘要。这里我们主要分析“符号检验”结果，其他检验的分析类似。

由图 5-30 可知，因为显著性水平为 0.05，而显著性概率 $P = 0.021 < 0.05$ ，所以拒绝原假设，可以得知新旧工艺之间差异的中位数不等于 0，即新旧工艺的成本有显著性差异。

双击输出文件中如图 5-30 所示的“假设检验摘要”表中“相关样本符号检验”行，打开如图 5-31 所示的模型浏览器，可以看到更详细的信息，其中检验统计量值等于 13.000；标准误差为 2.000；标准化检验统计量为 2.250，双侧渐进显著性为 0.024，小于 0.05，说明新旧工艺在成本上有显著性差异；精确显著性为 0.021，小于 0.05，也说明新旧工艺在成本上有显著性差异。



1. 由于个案数不超过 25 个，因此精确 p 值根据二项分布进行计算。

图 5-31 相关分析检验结果的模型浏览器

5.5 典型案例

5.5.1 判断某产品的需求量是否服从泊松分布

某产品在过去 200 个营业日里每天的需求量如表 5.11 所示，试在显著性水平 $\alpha = 0.05$ 的要求下，判断该产品的日需求量是否服从泊松分布。（数据来源：耿修林，《应用统计学》，科学出版社；参考数据文件：data5-10.sav。）

案例分析：本题最简单的方法是使用单样本 K-S 检验，也可以用卡方检验做分布的拟合优度检验，但卡方检验需要先获取泊松分布的期望频数。

表 5.11 某产品的日需求量

日需求量	0	1	2	3	4	5	6	7	8	9	10
天数	11	28	43	47	32	27	7	1	2	1	1

由于泊松分布的期望频数未知，下面我们介绍一种用 SPSS 来获取其泊松分布的期望概率值的方法：

- （1）将数据分成两列，一列是日需求量，其变量名为“demand”；另一列是天数，变量名为“days”，输入数据并保存，将变量“days”定义为权变量。
- （2）利用描述统计分析得到样本的均值为 3，该值作为泊松分布参数 $\hat{\lambda}$ 的无偏估计值。
- （3）执行“转换→计算变量”弹出如图 5-32 所示的“计算变量”对话框，在“目标变量”框中输入 p （即建立一个新变量 p ），在“数字表达式”框中输入泊松分布的求解函数：CDF.POISSON(demand, 3) - CDF.POISSON(demand - 1, 3)，单击“确定”按钮之后，变量 p 中的值就是日需求量 0-1 的期望概率值，分别为 0.0498、0.1494、0.2240、0.2240、0.1680、0.1008、0.0504、0.0216、0.0081、0.0008。
- （4）参照例 5-2 利用卡方检验可得到产品的需求量是否服从泊松分布。



图 5-32 “计算变量”对话框

5.5.2 调控政策前后大中城市住宅销售价格指数差异性分析

中国房价连连攀升，为了抑制房价，2011 年 1 月 26 日国务院公布了八条最新楼市调控政策，为了研究国八条对房价产生的影响，现抽样了 20 个大中城市在国八条公布前后两个月（2011 年 1 月及 6 月）的新建商品住宅价格同比增长指数，见表 5.12，检验公布国八条进行调控前后，大中城市住宅销售价格指数是否有显著差异。（数据来源：中华人民共和国国家统计局统计数据；参见数据文件：data5-11.sav。）

案例分析：因为本案例研究的是同一批对象实施国八条前后的房价同比增长指数，在数据组织时，一旦某个城市的 1 月同比增长指数位置确定，该城市对应的 6 月同比增长指数位置就固定了，不能再随意变动。所以应用探索性分析或单样本 K-S 检验对样本中的两组数据进行正态分布检验，若两组数据均为正态分布，则可以采用配对样本 T 检验进行分析；若样本中的两组数据不是正态分布，则应采用相关样本非参数检验进行分析。

表 5.12 20 个大中城市 2011 年 1 月及 6 月新建商品住宅价格同比增长指数

city	one	six	city	one	six
北京	109.1	102.7	南昌	108.9	108.4
天津	107.6	104.4	青岛	106	104.8
沈阳	109.8	106.8	郑州	109.6	106.6
大连	106.6	106	武汉	107.4	103.4
哈尔滨	107.6	104.5	长沙	110	108.3
上海	101.8	102.6	广州	100.1	105.4
南京	104.2	101.2	深圳	103.1	104.7
杭州	100.9	99.3	海口	121.6	100.7
合肥	106.5	100.7	重庆	108.1	106
厦门	105.1	106.7	成都	105	103.6

5.5.3 某行业企业赢利比例判断

对于一个行业，行业中企业的赢利比例是一个非常重要的指标，如果一个行业中大多数企业都能赢利，即企业赢利比例大，则说明这个行业发展性好，是一个朝阳行业；相反，如果企业赢利比例小，则说明行业前景堪忧。据统计，某地区某行业调查的 505 家企业中赢利的有 382 家，亏损的有 123 家，检验该行业企业赢利比例是否低于 0.8。数据组织及具体数据如表 5.13 所示。

表 5.13 某行业企业赢利与亏损统计

赢利类型 (type)	企业个数 (num)
1	382
2	123

（数据来源：夏怡凡，《SPSS 统计分析精要与实例详解》，电子工业出版社；参见数据文件：data5-12.sav。）

案例分析：因为本案例中企业赢利类型的取值只有两个：1 和 2，题目中要求企业赢利比例，即求赢利类型值为 1 的概率，所以本案例应采用二项分布检验进行分析。

5.5.4 棉条棉结杂质粒数分析

为了检验某纺织厂的生产情况是否正常，对该纺织厂连续 15 天生产的 28 号梳棉棉条的棉结杂质粒数进行了监控，数据中含有 15 个观测样本，代表 15 天，有 2 个属性变量：id（天数编号）、x1（棉结杂质粒数），具体数据如表 5.14 所示。（数据来源：杨维忠 等，《SPSS 统计分析 with 行业应用案例详解》，清华大学出版社；参见数据文件：data5-13.sav。）

案例分析：本案例要检验的是某棉纺厂生产情况正常与否，如果棉纺厂生产情况正常，它生

产的棉条的棉结杂质粒数应该还是比较稳定的，围绕某一个固定的常数在小范围内波动，而不应该是一组随机数。所以我们可以通过分析这组样本是否一组随机数来判断该棉纺厂生产情况是否正常。如果该组样本为一组随机数，说明该棉纺厂生产情况不正常；如果该组样本不是一组随机数，则认为该棉纺厂生产情况正常，所以本案例可以采用游程检验进行分析。

表 5.14 某企业连续 15 天棉结杂质粒数

天数编号	棉结杂质粒数（粒/g）	天数编号	棉结杂质粒数（粒/g）
001	71	009	77
002	69	010	69
003	68	011	68
004	75	012	64
005	74	013	70
006	67	014	63
007	70	015	61
008	76		

5.6 思考与练习

1. 掌握各种非参数检验的基本概念和思想，掌握它们的应用场合。
2. 独立样本和相关样本的区别以及它们的使用场合是什么？
3. 某地一周内各日忧郁症的人数分布如表 5.15 所示，请在显著性水平 0.05 下检验一周内各日忧郁症人数是否满足 1：1：2：2：1：1：1 的分布。（数据来源：余建英，《数据统计分析与 SPSS 应用》，人民邮电大学出版社；参见数据文件：data5-14.sav。）
4. 某厂质检部门对该厂的尼纶纤维进行检测，随机抽取 100 个样品，测得结果如表 5.16 所示，在显著性水平 $\alpha = 0.01$ 下，试判断尼纶纤维度是否与正态分布相吻合。（数据来源：耿修林，《应用统计学》，科学出版社；参见数据文件：data5-15.sav。）

表 5.15 忧郁症人数分布表

星期	1	2	3	4	5	6	7
患者数	31	38	70	80	29	24	31

表 5.16 尼纶纤维度数据

纤维度	1.28	1.31	1.34	1.37	1.4	1.43	1.46	1.49	1.52	1.55
频数	1	4	7	22	23	25	10	6	1	1

5. 为了研究紧张对人的影响，实验者让 18 位大学生用两种方法打同样的结。其中一半受试者先学 A 方法，后学 B 方法；另一半先学 B 方法，后学 A 方法。在一天夜里，突然要求每个受试者打这样的结。结果选择先学方法的有 16 人，选择后学方法的有 2 人。在显著性水平 $\alpha = 0.05$ 下检验紧张时用先学方法打结的概率和用后学方法打结的概率是否有显著性差异。（参见数据文件：data5-16.sav。）
6. 投掷一枚硬币 30 次，得到由正面（H）和反面（T）组成的序列如下：

HTTHTHHHTHTHTTHTHTHTHTHTTHTHTHTHT

请在显著性水平 $\alpha = 0.05$ 下检验此序列是否随机。（数据来源：M.R.斯皮格尔，《统计学（第 3 版）》，科学出版社；参见数据文件：data5-17.sav。）
7. 某农民想了解两品种的小麦 I、II 产量是否有显著区别，其产量数据如表 5.17 所示，分别在显著性水平 0.05 和 0.01 下检验两品种产量是否有显著性差异。（数据来源：M.R.斯皮格尔，《统计学（第 3 版）》，科学出版社；参见数据文件：data5-18.sav。）

表 5.17 两种小麦的产量数据

小麦 1	15.9	15.3	16.4	14.9	15.3	16	14.6	15.3	14.5	16.6	16
小麦 2	16.4	16.8	17.1	16.9	18	16	18.1	17.2	15.4		

8. 为研究长跑运动对增强普通高校学生心脏功能的效果，对某校 15 名男生进行测试，经过 5 个月的长跑锻炼后看其晨脉是否减少。锻炼前后的晨脉数据如表 5.18 所示。

表 5.18 长跑锻炼前后晨脉变化表

锻炼前	70	76	56	63	63	56	58	60	65	65	75	66	56	59	70
锻炼后	48	54	60	64	48	55	54	45	51	48	56	48	64	50	54

试问锻炼前后的晨脉在显著性水平 0.05 下有无显著性差异。（数据来源：卢纹岱，《SPSS for Windows 统计分析（第 3 版）》，电子工业出版社；参见数据文件：data5-19.sav。）

9. 某公司培训部门为了解四种训练员工方案的效果，决定将新招收的 30 名大学应届毕业生随机分成 4 个组，分别按不同的训练方案进行培训，训练结束后进行测试，所得数据如表 5.19 所示，在显著性水平 0.05 和 0.01 下检验训练方案之间是否有显著性差异。（数据来源：耿修林，《应用统计学》，科学出版社；参见数据文件：data5-20.sav。）

表 5.19 30 名受试者的测试成绩

A	66	74	82	75	73	97	87	
B	72	51	59	62	74	64	78	
C	61	60	57	60	81	55	70	71
D	63	61	76	84	58	65	69	80

10. 使用 4 种不同的容器存放果汁，经过半年的存放以后，请 8 位品尝员品尝，每位品尝员给这 4 种容器的果汁味道打分，得到的数据如表 5.20 所示，在显著性水平 0.05 下检验存放果汁的容器之间是否有显著性差异。（数据来源：郝黎仁，《SPSS 实用统计分析》，中国水利水电出版社；参见数据文件：data5-21.sav。）

表 5.20 四种容器的果汁味道分数表

人员	容器 1	容器 2	容器 3	容器 4
1	4.81	5.54	6.55	6.14
2	5.09	5.61	6.29	5.72
3	6.61	6.6	7.4	6.9
4	5.03	5.7	6.4	5.8
5	5.15	5.31	6.28	6.23
6	5.05	5.58	6.26	6.06
7	5.77	5.57	6.22	5.42
8	6.17	5.84	6.76	6.04

11. 假如有 100 个摇奖数据，全部数据在表 5.21 中给出，试分析这些摇奖数据是否均匀分布在 0~9。（参见数据文件：data5-22.sav。）

表 5.21 100 个摇奖数据

483883578678987972779790937037967599927771341595797045224821697137
9558447985127910418848851571922876

第6章 方差分析

第4章介绍的T检验是为了解决两样本间均值比较的问题，但在研究中经常遇到两个以上均值比较的问题。例如，农作物的种植过程中，产量会受品种、施肥量、土地质量等众多因素的影响，而每种因素的影响大小是不同的，因此找到其中的关键（重要）因素就很重要。进一步，在掌握了关键因素，如品种、施肥量等以后，还需对不同的品种、不同的施肥量进行对比分析，研究究竟哪个品种的产量高，施肥量多少合适，哪种品种与哪种施肥水平搭配最优秀。对于多个总体两两进行独立样本T检验是一种处理方法，但随着总体数目增多，这种方法的弊端会越来越明显，这就需要引入另一种统计分析方法——方差分析。

本章主要介绍方差分析的基本概念和常用的方差分析方法：单因素方差分析、多因素方差分析、协方差分析和多元方差分析。

6.1 方差分析简介

6.1.1 方差分析的概念

方差分析（Analysis of Variance）最早是由英国统计学家 R.A.Fisher 于 1920 年前后对农业试验作统计分析时提出来的。由于它可以由较少的试验有效地获得大量的信息，所以已广泛应用于工业、商业、生物、医学等众多领域。

方差分析中的常用术语如下。

- 观测变量：也叫因变量，如上例中的作物产量。
- 控制变量：影响实验结果的自变量，也称因子，如上例中的品种、施肥量等。
- 水平：控制变量的不同类别，如 A 品种，B 品种；10 公斤化肥、20 公斤化肥、30 公斤化肥等。
- 随机因素：因素的水平与实验结果的关系是随机的，即不确定因素。

由于受考察因素以及各种随机因素的影响，试验所得的数据呈现波动状。造成波动的原因可分为两类，一类是试验中施加的对观测变量形成影响的控制变量，另一类是不可控制的随机因素。方差分析认为，如果控制变量的不同水平对观测变量产生了显著影响，那么，它和随机变量共同作用必然使得观测变量值有显著变动；反之观测变量值的变动就不明显，其变动可以归结为随机因素的变动造成的。进一步，如何判断控制变量在不同水平上的观测变量值是否产生了显著波动呢？判断的原则是，如果在控制变量各水平下，观测变量总体的分布出现了显著差异，则认为观测变量值发生了明显的波动，意味着控制变量的不同水平对观测变量产生了显著影响；反之，观测变量总体的分布没有出现显著差异，则认为观测变量值没有发生显著波动，说明控制变量的不同水平对观测变量没有产生显著影响。

方差分析对观测变量各总体的分布还有两个基本假设：① 观测变量各总体应服从正态分布 $N(\mu_i, \sigma_i^2)$ ；② 观测变量总体的方差应相等，即方差具有齐性： $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$ 。基于上述两个基本假设，方差分析对各总体分布是否有显著差异的推断就转化为对各总体均值是否存在显著差异的推断了。

根据控制变量的个数可以将方差分析分成单因素方差分析、多因素方差分析及协方差分析。根据观测变量的个数可以分为一元方差分析和多元方差分析。

6.1.2 方差分析的一般步骤

方差分析通常按如下基本步骤进行：

第 1 步 方差分析条件检测。

方差分析条件即前面提出的两个基本假设：服从正态分布和方差齐性，还有一个就是控制变量的类别（即水平数量）有限，用 SPSS 术语来说就是控制变量是取值有限的名义尺度或顺序尺度变量。

第 2 步 提出原假设。

控制变量在不同水平下观测变量各总体均值无显著差异，对应协方差分析，则是扣除协变量影响后，控制变量不同水平下观测变量各总体均值无显著差异。

第 3 步 构造检验的统计量。

针对不同分析方法的数学模型，计算平方和、均方，并计算检验统计量（ F ）。

第 4 步 统计决策。

如果 F 值对应的显著性概率 P 值小于给定显著性水平 α ，则拒绝原假设，认为控制变量不同水平下各总体均值有显著差异；反之，认为控制变量不同水平下各总体均值没有显著差异。

6.2 单因素方差分析

6.2.1 基本概念及统计原理

1. 基本概念

单因素方差分析用来研究一个控制变量的不同水平是否给观测变量造成了显著差异和变动。例如：培训是否给学生成绩造成了显著影响；不同学历是否对工资收入造成了影响；不同地区考生的成绩是否有显著差异等。

2. 统计原理

采用的统计推断方法是计算 F 统计量，进行 F 检验。总的变异平方和记为 SST，分解为两部分：一部分是由控制变量引起的离差，记为 SSA（组间 Between Groups 离差平方和）；另一部分是由随机变量引起的离差，记为 SSE（组内 Within Groups 离差平方和）。于是有

$$SST = SSA + SSE \tag{6.1}$$

式中，

$$SSA = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 \tag{6.2}$$

式中， k 为水平数； n_i 为第 i 个水平下的样本容量； \bar{x}_i 为控制变量第 i 个水平下观测变量的样本均值； \bar{x} 为观测变量的均值。可见，组间样本离差平方和是各水平组均值和总体均值离差的平方和，反映了控制变量的不同水平对观测变量的影响。

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \tag{6.3}$$

式中， x_{ij} 为控制变量在第 i 水平下的第 j 个样本值。组内离差平方和是每个数据与本水平组平均值离差的平方和，反映了数据抽样误差的大小程度。

F 统计量是平均组间平方和与平均组内平方和的比值，即

$$F = \frac{SSA / (k - 1)}{SSE / (n - k)}$$

(6.4)

从 F 值的计算公式可以看出，如果控制变量的不同水平对观测变量有显著影响，那么观测变量的组间离差平方和就大， F 值也就较大；反之，如果控制变量的不同水平没有对观测变量造成显著影响，那么组内离差平方和的影响就会比较小， F 值就比较小。

3. 分析步骤

方差分析问题属于统计推断中的假设检验问题，其基本步骤与假设检验完全一致，具体如下。

第 1 步 提出原假设。

单因素方差分析的原假设 H_0 ：控制变量不同水平下观测变量各总体均值无显著差异，控制变量不同水平下的效应同时为 0，记为： $\mu_1 = \mu_2 = \cdots = \mu_k$ ，即说明控制变量的不同水平对观测变量没有产生显著性影响。

第 2 步 选择检验统计量。

方差分析采用的是 F 统计量，计算公式如式 (6.4)，服从 $(k-1, n-k)$ 个自由度的 F 分布。

第 3 步 计算检验统计量的观测值和概率 P 值。

SPSS 会根据式 (6.4) 自动计算 F 统计值，并依据 F 分布表给出相应的显著性概率值 P 。不难理解，如果控制变量对观测变量造成了显著影响，观测变量总的变差中控制变量影响所造成的比例相对于随机变量就会较大， F 值显著大于 1；反之，观测变量的变差应归结为随机变量造成的， F 值接近于 1。

第 4 步 给出显著性水平 α ，作出决策。

如果显著性概率 P 值小于显著性水平 α ，则拒绝原假设，认为控制变量不同水平下各总体均值有显著差异；反之，认为控制变量不同水平下各总体均值没有显著差异。

6.2.2 单因素方差分析 SPSS 实例分析

【例 6-1】用四种饲料喂猪，共 19 头，分为四组，每一组用一种饲料。一段时间后称重，猪的体重增加数据如表 6.1 所示，比较四种饲料对猪体重增加的作用有无不同。（数据来源：金丕焕，《医用统计方法》，人民卫生出版社；参见数据文件：data6-1.sav.）

第 1 步 分析。

由于考虑的是一个控制变量（饲料）对一个观测变量（猪的体重）的影响，而且是 4 种饲料，所以不适宜用独立样本 T 检验（仅适用两组数据），应采用单因素方差分析。

第 2 步 数据的组织。

数据分成两列，一列是猪的体重，变量名为“weight”，另一变量是饲料品种（变量值分别为 1，2，3，4），变量名为“fodder”，并将标签设成对应的中文名称，输入数据并保存。

第 3 步 方差相等的齐性检验。

由于方差分析的前提是各个水平下（这里是不同的饲料 fodder 影响下的体重 weight）的总

表 6.1 饲料比较数据资料

饲料 A	饲料 B	饲料 C	饲料 D
133.8	151.2	193.4	225.8
125.3	149.0	185.3	224.6
143.1	162.7	182.8	220.4
128.9	143.8	188.5	212.3
135.7	153.5	198.6	

体服从方差相等的正态分布，且各组方差具有齐性。其中正态分布的要求并不是很严格，但对于方差相等的要求是比较严格的，因此必须对方差相等的前提进行检验。选择菜单：“分析→比较平均值→单因素 ANOVA 检验”，打开如图 6-1 所示的“单因素 ANOVA 检验”对话框。

该对话框主要由以下几部分组成。

(1) 候选变量框：即左侧变量列表框。

(2) “因变量列表”框：选择单因素方差分析的观测变量，可以同时选择多个变量，此时 SPSS 就将分别对各观测变量作单因素方差分析。本例中我们选择“猪重 (weight)”变量。

(3) “因子”框：选择因素变量（也称控制变量），由于进行的是单因素方差分析，所以此时只能选择一个因变量。本例中我们选择“饲料品种 (fodder)”变量。

(4) “对比”按钮：单击“对比 (N) ...”按钮，弹出如图 6-2 所示的“单因素 ANOVA 检验：对比”对话框。



图 6-1 “单因素 ANOVA 检验”对话框



图 6-2 “单因素 ANOVA 检验：对比”对话框

该对话框主要用于对组间平方和划分趋势成分，或者指定先验对比。主要包括如下几项。

- “多项式”复选框：选择是否对方差分析的组间平方和进行分解并进行趋势检验。
- “等级”下拉列表：选中“多项式”复选框之后，该下拉列表被激活，用于选择进行趋势检验的曲线类型，主要有线性、二次、立方、四次、五次多项式。如果选择了高次方曲线，系统会给出所有相应各低次方曲线的拟合优度检验结果以供选择。
- “系数”框：精确定义某些组间平均数的比较。一般按照分组变量顺序给每组一个系数值，但所有系数值之和为 0。列表中第一个系数对应于分类变量的最小值，最后一个系数对应于最大值。输入方法是在“系数”文本框中输入一个系数，单击“添加”按钮，使之进入下面的列表框中。因变量分几组，就输入几个系数，多出的无意义。

(5) “事后比较”按钮：单击图 6-1 右边“事后比较”按钮，打开如图 6-4 所示进行事后比较选项设置的对话框，将在第 4 步介绍。

(6) “选项”按钮：单击图 6-1 右边“选项 (O) ...”按钮，弹出如图 6-3 所示的“单因素 ANOVA 检验：选项”对话框，本例选择“方差齐性检验”。

该对话框主要包括如下选项。

① “统计”复选框组：可以选择需要输出的统计量，主要有以下几种。

- 描述：即要求输出描述统计量，包括观测量数目、均值、



图 6-3 “单因素 ANOVA 检验：选项”对话框

最小值、最大值、标准差、标准误差以及各组中每个因变量的 95%置信区间。

- 固定和随机效应：输出不变效应模型和随机效应模型的标准差、标准误差以及 95%置信区间。
- 方差齐性检验：表示要求用莱文（Levene）统计量进行方差一致性检验，该方法不依赖于正态假设，即不要求样本一定服从正态分布。
- 布朗-福塞斯：表示计算分组均数相等的布朗-福塞斯（Brown-Forsythe）统计量，当不能把握方差齐性假设时，此统计量比 F 统计量更优。
- 韦尔奇：表示计算分组均数相等的韦尔奇（Welch）统计量，当不能把握方差齐性假设时，此统计量比 F 统计量更优。

② “平均值图”复选框：选中该复选框，表示输出均数分布图，根据因素变量值所确定的各组均数描绘出因变量的均值分布情况。

③ “缺失值”处理选项组：选择缺失值的处置方式。

设置好选项后，运行结果如下。

表 6.2 不同饲料的方差齐性检验结果（猪重）

莱文统计	自由度 1	自由度 2	显著性
.024	3	15	.995

表 6.3 几种饲料的方差检验（ANOVA）结果（猪重）

	平方和	自由度	均方	F	显著性
组间	20538.698	3	6846.233	157.467	.000
组内	652.159	15	43.477		
总计	21190.858	18			

方差齐性检验的 H_0 假设是：方差具有齐性。从表 6.2 可看出显著性概率 P 值 = 0.995 > α (0.05)，说明应该接受 H_0 假设（即方差具有齐性）。故下面就用方差具有齐性的检验方法。当然表 6.3 就是几种饲料方差分析的结果，组间平方和为 20538.698，自由度（df）为 3，均方为 6846.233；组内平方和为 652.159，自由度为 15，均方为 43.477； F 统计量为 157.467。由于组间比较的显著性概率 $P = 0.000 < 0.05$ ，故应拒绝 $H_0 =$ 假设（四种饲料喂猪效果无显著差异），说明四种饲料对养猪的效果有显著性差异。

第 4 步 多重比较分析。

通过上面的步骤，判断出了 4 种饲料喂猪效果有显著性差异。如果想进一步了解究竟是哪种饲料与其他组有显著性的均值差别（即哪种饲料更好）等细节问题，就需要在多个样本均值间进行两两比较。在图 6-1 中单击“事后比较（H）...”按钮，如图 6-4 所示。

该对话框主要用于定义多重比较的检验方法。比如，方差分析的结果认为因素 A 各水平之间的差异会对观测变量 X 造成显著影响，但这并不意味着任意两个水平之间的差异都会对 X 造成显著影响。要解决这个问题，就有必要将各水平的均值进行两两比较，这种两两比较的方法就称为多重比较。该对话框各项意义如下。

① “假定等方差”选项组，该选项组给出方差相等时的多重比较方法，具体有 14 种，



图 6-4 “单因素 ANOVA 检验：事后多重比较”选项框

其中常用的方法有如下 6 种。

- LSD: Least-Significant Difference 检验法。用 T 检验完成各组间的配对比较, 检验的敏感性高, 各水平间的均值存在的微小差异也有可能被检验出来, 但此方法对第一类弃真错误的概率不进行调整。α 可指定为 0~1 之间任何显著性水平, 默认值为 0.05。
- S-N-K: Student-Newman-Keuls 检验法, 用 Student-Range 分布进行各组均值间的配对比较。如果各组样本含量相等或者选择了 Harmonic average of all groups (各组样本含量的调和平均数), 即用各组样本含量的调和平均数进行样本量估计, 还将用逐步过程进行齐次子集 (差异较小的子集) 的均值配对比较。在该过程中各组均值按从大到小的顺序排列, 最先比较最极端的差异。α 只能取 0.05。
- 邦弗伦尼: Bonferron 检验法, 即修正最小显著差异。用 T 检验完成各组间的配对比较, 同时通过设置每个检验的误差率来控制第一类错误的概率。α 可指定为 0~1 之间任何显著性水平, 默认值为 0.05。
- 邓肯: Duncan 检验法, 指定一系列 Range 值, 逐步计算、比较, 得出结论。α 可取 0.01、0.05 和 0.1, 默认值为 0.05。
- 图基: Tukey's honestly significant difference 方法, 即 Tukey 显著差异法。用 Student-Range 统计量进行所有组间均值的配对比较, 用所有配对比较误差率作为实验误差率。α 只能取 0.05。
- 雪费: Scheffe 差别检验法。使用 F 统计量作为检验统计量, 对所有可能的组合进行同步的配对比较, 可用于检验分组均值所有可能的线性组合, 但灵敏度不太高。α 可取 0~1 之间任何显著性水平, 默认值为 0.05。

② “不假定等方差”选项组, 该选项组给出方差不相等时的确定多重比较方法, 包括如下 4 项, 其中, 邓尼特 (Dunnett's C) 方法较常用。

- 塔姆黑尼 T2: Tamhane's T2 检验法。用 T 检验进行各组均值配对比较。
- 邓尼特 T3: Dunnett's T3 检验法。用基于 Student 最大模数的比较配对试验。
- 盖姆斯-豪厄尔: Games-Howell 检验法。指方差不齐时的配对比较检验, 该方法较灵活。
- 邓尼特 C: Dunnett's C 检验法。用 Student-Range 极差统计量进行配对比较检验。

③ “显著性水平”文本框: 设定各种多重比较检验的显著性水平, 系统默认值 α 为 0.05, α 可取的值一般为 0.01、0.05 和 0.1。

多重比较检验法以矩阵的形式输出检验结果, 在确定的显著性水平下, 对那些组均值有显著差异的分组用 “*” 标记出来。

由于第 3 步检验出方差具有齐性, 故选择一种方差相等的方法, 这里选 LSD 方法; 显著性水平取 0.05。之后, 返回图 6-1, 单击 “选项 (O) ...” 按钮, 弹出如图 6-3 所示的 “选项” 对话框, 选中 “描述” 和 “平均值图”, 对数据进行整体描述并绘制几种饲料作用下猪重的均数分布图。

第 5 步 运行结果及分析。

完成以上操作步骤后, 生成表 6.4、表 6.5 及图 6-5 所示的结果, 具体意义分析如下。

表 6.4 描述性统计表 (猪重)

	个案数	平均值	标准差	标准误差	平均值的 95% 置信区间		最小值	最大值
					下限	上限		
1	5	133.3600	6.80794	3.04460	124.9068	141.8132	125.30	143.10
2	5	152.0400	6.95723	3.11137	143.4015	160.6785	143.80	162.70
3	5	189.7200	6.35035	2.83996	181.8350	197.6050	182.80	198.60
4	4	220.7750	6.10594	3.05297	211.0591	230.4909	212.30	225.80
总计	19	171.5105	34.31137	7.87157	154.9730	188.0481	125.30	225.80

(1) 描述性统计表

表 6.4 为描述性统计量结果，给出了四种饲料分组的样本含量 N 、观测变量猪体重的平均值 (Mean)、标准差 (Std.Deviation)、标准误差 (Std.Error)、平均值的 95% 置信区间、最小值和最大值。

(2) 多重比较结果表

表 6.5 是方差分析的多重比较结果，分别进行了饲料品种的两两比较，以第 1 种饲料与第 2、3、4 种的比较为例，对猪重影响的均值分别相差 18.68000、56.36000 和 87.41500，而且所有的显著性概率 P 值 = 0.000 < 0.05，这说明第 1 种饲料与其他三种饲料均具有显著性差异，而且体重均值均低于其他 3 种饲料，这说明第 1 种饲料的效果没有和其他三种饲料好。“*”表示不同饲料之间存在显著性差异。整个表反映出来四种饲料之间均存在显著性差异，从效果来看第 4 种最好，其次是第 3 种，第 1 种最差。

表 6.5 多重比较结果表（猪重）（LSD）

(I) 饲料品种	(J) 饲料品种	平均值差值 (I-J)	标准误差	显著性	95% 置信区间	
					下限	上限
1	2	-18.68000*	4.17024	.000	-27.5687	-9.7913
	3	-56.36000*	4.17024	.000	-65.2487	-47.4713
	4	-87.41500*	4.42321	.000	-96.8428	-77.9872
2	1	18.68000*	4.17024	.000	9.7913	27.5687
	3	-37.68000*	4.17024	.000	-46.5687	-28.7913
	4	-68.73500*	4.42321	.000	-78.1628	-59.3072
3	1	56.36000*	4.17024	.000	47.4713	65.2487
	2	37.68000*	4.17024	.000	28.7913	46.5687
	4	-31.05500*	4.42321	.000	-40.4828	-21.6272
4	1	87.41500*	4.42321	.000	77.9872	96.8428
	2	68.73500*	4.42321	.000	59.3072	78.1628
	3	31.05500*	4.42321	.000	21.6272	40.4828

*. 平均值差值的显著性水平为 0.05。

(3) 均值折线图

图 6-5 为几种饲料作用下的猪重均值的折线图，可以看出均值分布比较陡峭，均值差异也较大。

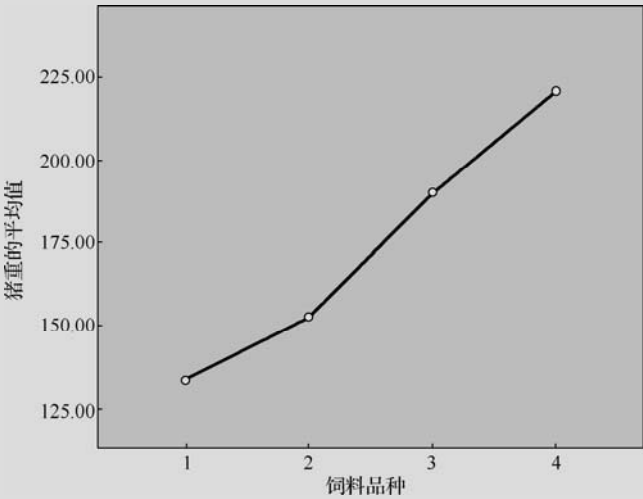


图 6-5 几种饲料作用下的猪重均值折线图

☆说明☆

- (1) 在方差分析中，要求每一组资料服从正态分布，并不是要求各组资料服从一个正态分布（因为这就意味着各组的总体均值相同，失去统计检验的必要性），所以不能把各组的资料合在一起做正态性检验。总的来讲，方差分析对正态性具有稳健性，即偏态分布对方差分析的结果影响不会太大，样本量较大时，方差分析对正态性要求大大降低（根据中心极限定理可知：大样本均数近似服从正态分布）。并且由于大多数情况下，样本资料只是近似服从正态分布而不是完全服从正态分布。在大样本情况下，用正态性检验只能检验是否服从正态分布，而不能检验是否近似正态分布。故在大样本情况下，考察资料的近似正态性，应用频数图，当然在大多数情况下也可不做正态性检验。
- (2) 前面讲了方差齐性是方差分析的前提条件，为何又有“未假定方差齐性”下的几种分析方法呢？原因是：方差齐性的假定是在最小二乘估计的框架下谈的，在广义最小二乘法的框架下面，已经不是方差分析的条件。我们从 SPSS 8.0 开始把菜单名字转换为一般线性模型就可以看出这种变化，即 SPSS 已经把方差分析放在了广义最小二乘法的框架下。其实在多数情况下用两类方法分析的结果是一致的。
- (3) 当然，也可将方差齐性检验和多重比较一起进行，只不过在不知道方差是否具有齐性的情况下，可分别同时选上一两种方差具有齐性和方差不具有齐性的多重比较方法，分析结果出来后，从方差齐性分析表分析方差是否一致，再选择相应的分析方法。例如，上例可同时选上 LSD 和邓尼特 C 两种方法，当分析出方差具有齐性时，选择 LSD 的分析结果，方差不一致时选择邓尼特 C 的分析结果。

6.3 多因素方差分析

6.3.1 基本概念及统计原理

多因素方差分析用来研究两个及两个以上的控制变量是否对观测变量产生显著性影响。由于讨论多个因素对观测变量的影响，因此这种方差分析过程称为多因素方差分析。多因素方差分析不仅能够分析多个控制因素对观测变量的影响，也能够分析多个控制因素的交互作用对观测变量产生的影响，进而最终找到利于观测变量的最优组合。

1. 基本概念

例如，分析不同品种、不同施肥量对农作物产量的影响时，可将农作物产量作为观测变量，品种和施肥量作为控制变量。利用多因素分析方法，研究不同品种、不同施肥量是如何影响农作物产量的，并进一步研究哪种品种与哪种水平的施肥量是提高农作物产量的最优组合。

多因素方差分析不仅需要分析多个控制变量独立作用对观测变量的影响，还要分析多个控制变量的交互作用对观测变量的影响，及其他随机变量对结果的影响。因此，需要将观测变量总的离差平方各分解为 3 个部分：

- (1) 多个控制变量单独作用引起的离差平方和。
- (2) 多个控制变量交互作用引起的离差平方和。
- (3) 其他随机因素引起的离差平方和。

2. 统计原理

以两个控制变量为例，多因素方差分析将观测变量的总离差平方和分解为

$$SST=SSA+SSB+SSAB+SSE \quad (6.5)$$

式中， SST 为观测变量的总离差平方和； SSA 、 SSB 分别为控制变量 A 、 B 独立作用的离差平方和； $SSAB$ 为控制变量 A 和 B 交互作用引起的离差平方和； SSE 为随机变量引起的误差。通常称 $SSA+SSB$ 为主效应(Main Effects)， $SSAB$ 为多向交互效应(N-Way Effects)， SSE 为剩余(Residual)。

在两因素方差分析中， SST 的定义同式 (6.5)。设控制变量 A 有 k 个水平，变量 B 有 r 个水平。 SSA 的定义为

$$SSA = \sum_{i=1}^k \sum_{j=1}^r n_{ij} (\bar{x}_i^A - \bar{x})^2 \quad (6.6)$$

式中， n_{ij} 为因素 A 第 i 个水平和因素 B 第 j 个水平下的样本观测值个数， \bar{x}_i^A 为因素 A 第 i 个水平下观测变量的均值。 SSB 的定义与 SSA 的定义类似。 SSE 定义为

$$SSE = \sum_{i=1}^k \sum_{j=1}^r \sum_{l=1}^{n_{ij}} (x_{ijl} - \bar{x}_{ij}^{AB})^2 \quad (6.7)$$

式中， \bar{x}_{ij}^{AB} 是因素 A 、 B 在水平 i 、 j 下的观测变量均值。于是交互作用可解释的离差为

$$SSAB = SST - SSA - SSB - SSE \quad (6.8)$$

在多因素方差分析中，控制变量可进一步划分为固定效应模型和随机效应模型。其中固定效应通常指控制变量的各水平是可以严格控制的，它们对观测变量带来的影响是固定的，如温度、品种等；随机效应是指控制变量的各水平无法作严格的控制，它们对观测变量带来的影响是随机的，如城市规模、受教育水平等。两种效应的主要差别体现在统计量的构造上。

在固定效应模型中，各 F 统计量为

$$F_B = \frac{SSB / (r-1)}{SSE / kr(l-1)} = \frac{MSB}{MSE} \quad (6.9)$$

$$F_A = \frac{SSA / (k-1)}{SSE / kr(l-1)} = \frac{MSA}{MSE} \quad (6.10)$$

$$F_{AB} = \frac{SSAB / (k-1)(r-1)}{SSE / kr(l-1)} = \frac{MSAB}{MSE} \quad (6.11)$$

在随机效应模型中， F_{AB} 统计量不变，其他两个 F 统计量分别为

$$F_A = \frac{SSA / (k-1)}{SSAB / (k-1)(l-1)} = \frac{MSA}{MSAB} \quad (6.12)$$

$$F_B = \frac{SSB / (r-1)}{SSAB / (k-1)(l-1)} = \frac{MSB}{MSAB} \quad (6.13)$$

3. 分析步骤

多因素方差分析问题亦属于统计推断中的假设检验问题，其基本步骤与假设检验一致，具体如下。

第 1 步 提出原假设。

多因素方差分析的原假设 H_0 : 各控制变量不同水平下观测变量各总体均值无显著性差异, 控制变量各效应和交互作用效应同时为 0, 即控制变量和它们的交互作用对观测变量没有产生显著性影响。数学表达式为 $a_1=a_2=\cdots=a_k=0, b_1=b_2=\cdots=b_r=0$ 。

第 2 步 构造检验统计量。

多因素方差分析采用的是 F 统计量, 根据效应模型选择情况, 计算公式如式 (6.9) ~ (6.13)。

第 3 步 计算检验统计量的观测值和概率 P 值。

SPSS 会自动将相关数据代入各式, 计算出检验统计量的观测值的显著性概率 P 值 (也称相伴概率值)。

第 4 步 给出显著性水平 α , 作出决策。

给定显著性水平 α (系统默认为 0.05), 并与各个检验统计量的概率 P 值进行比较。在固定效应模型中, 如果 F_A 的概率 P 值小于显著性水平 α , 则应拒绝原假设, 认为控制变量 A 的不同水平对观测变量的均值产生了显著性影响; 反之, 则应接受原假设, 认为控制变量 A 的不同水平对观测变量没有产生显著性影响。同理, 可对 B 的显著性及 A 和 B 的交互作用的显著性作推断。

6.3.2 多因素方差分析 SPSS 实例分析

【例 6-2】 研究一个班三组不同性别的学生分别接受了三种不同的教学方法后, 在数学成绩上是否有显著性差异, 数据如表 6.6 所示。(参见数据文件: data6-2.sav。)

表 6.6 三组不同性别和不同教学方法学生的数学成绩

姓名	数学	组别	性别	姓名	数学	组别	性别
张青华	99	0	m	蔡春江	67	1	m
王洁云	88	0	f	武佳琪	56	1	f
吴凌风	99	0	m	陈雪吟	56	1	m
刘行	89	0	m	罗超波	79	2	m
马萌	94	0	f	尹珣	56	2	f
单玲玲	90	0	m	张敏	89	2	m
宋丽君	55	1	f	郭晓艳	99	2	m
辛瑞晶	50	1	m	李福利	70	2	f
王滢滢	67	1	f	罗帆	89	2	m

第 1 步 分析。

研究不同教学方法和不同性别对数学成绩的影响。这是一个多因素 (双因素) 方差分析问题。

第 2 步 数据组织。

按表 6.6 的变量名组织成 4 列数据, 数据文件为 data6-2.sav。

第 3 步 变量设置。

选择菜单“分析→一般线性模型→单变量”, 打开“单变量”对话框, 并按图 6-6 所示进行设置。

该对话框主要由以下几部分组成。

(1) 候选变量框: 即左侧变量列表框。

(2) “因变量”框: 选择多因素方差分析的观测变量, 从左侧的变量列表框中移入。只能选择一个而且是数值型的变量。此时 SPSS 将对所选的观测变量做多因素方差分析。

(3) “固定因子”框: 选择控制变量, 由于进行的是多因素方差分析, 所以可选择多个变量(数值型和字符串型均可)。

(4) “随机因子”框: 选择随机因素变量。

(5) “协变量”框: 选择协变量, 此功能将在下一节用到。

(6) “WLS 权重”框: 选择加权最小二乘法的权重系数的变量。

第4步 设置方差齐性检验。

单击“选项(O)…”按钮, 弹出“单变量: 选项”对话框, 并按图 6-7 所示设置。



图 6-6 “单变量”对话框

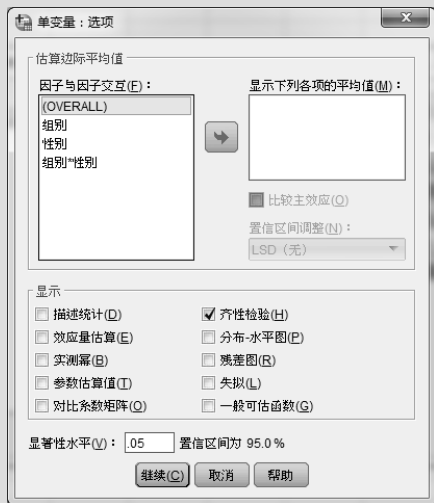


图 6-7 “单变量: 选项”对话框

对话框主要由以下几部分组成。

① “因子与因子交互”框: 列出可选的控制变量及其交互作用, 其中 OVERALL 表示对所有控制变量及其交互作用都计算其相应的样本均值。

② “显示下列各项的平均值”框: 将因子与因子交互框中要计算均值的变量选入此框中。

③ “比较主效应”复选框: 当显示均值框中有元素时, 该项被激活, 用来定义是否对选中变量进行均值的多重比较。

④ “置信区间调整”下拉列表: 选择多重比较的方法。

⑤ “显示”复选框组: 定义输出的统计量。

由于方差分析要求不同组别数据方差具有齐性, 故应进行方差齐性检验, 本例选中“齐性检验”, 显著性水平设为默认值 0.05。

第5步 设置控制变量的多重比较分析。

通过以上步骤只能判断两个控制变量的不同水平是否对观察变量产生了显著影响。如果想进一步了解究竟是哪个组与其他组有显著的均值差别, 就需要进行控制变量的多重比较分析(这与前面的单因素方差分析一致)。单击“事后比较(H)…”按钮, 弹出如图 6-8 所示的对话框, 在其中选出需要进行比较分析的控制变量, 这里选“组别”, 再选择一种方差相等时的检验模型, 如 LSD。

第6步 选择建立多因素方差分析的模型种类。

单击“模型...”按钮弹出如图 6-9 所示的对话框。



图 6-8 “单变量：实测平均值的事后多重比较”对话框



图 6-9 “单变量：模型”对话框

该对话框主要用来定义方差分析的模型，主要包括以下几部分。

- (1) “全因子”单选按钮：系统默认选项，包含所有因子主效应、所有协变量主效应以及所有因子间交互。它不包含协变量交互。
 - (2) “定制”单选按钮：指定其中一部分的交互或指定因子协变量交互，选择该项，则激活“类型”下拉列表，选择感兴趣的主体内效应和交互以及主体间效应和交互。
 - (3) “平方和”下拉列表：计算主体间模型平方和的方法，一般默认选择Ⅲ类。
- 本例用默认的全因子模型。

第 7 步 以图形方式展示交互效果。

如果各因素间无交互作用，则各个水平对应的图形应趋于平行，否则相交。单击“图(T)…”按钮，弹出“单变量：轮廓图”对话框，如图 6-10 所示，设置控制变量的交互效果，将“组别”和“性别”变量分别移入“水平轴”和“单独的线条”框后，单击“添加”按钮添加到下方的文本框中。

第8步 对控制变量各水平上的观察变量的差异进行对比检验。

单击“对比(T)…”按钮,弹出“单变量:对比”对话框,如图6-11所示,对两种因子水平进行对比分析,用“简单”方法,并以“最后一个”水平的观察变量均值为标准(选择“对比”方式后需单击“变化量”按钮进行确认)。



图 6-10 “单变量:轮廓图”对话框

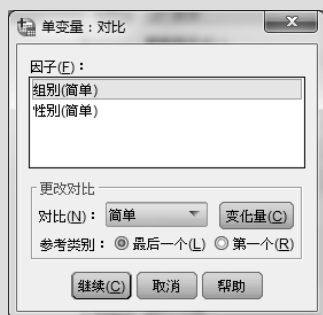


图 6-11 “单变量:对比”对话框

第9步 主要结果及分析。

完成以上设置后单击“确定”按钮运行,结果如表6.7~表6.12及图6-12所示,各图表的具体分析如下。

表 6.7 分组描述表主体间因子

		N
组别	0	6
	1	6
	2	6
性别	f	7
	m	11

(1) 分组描述

表6.7表示各控制变量的分组情况。

(2) 方差齐性检验结果

表6.8是方差齐性检验的计算结果。显著性概率 P 值 $=0.879>0.05$,因此可以认为各个组总体方差是相等的,满足方差检验的前提条件。

表 6.8 方差齐性检验结果 误差方差的莱文等同性检验^a 因变量: 数学

F	自由度 1	自由度 2	显著性
.339	5	12	.879

检验“各个组中的因变量误差方差相等”这一原假设。

a. 设计: 截距 + 组别 + 性别 + 组别 * 性别

(3) 多因素方差分析及交互检验结果

表6.9是多因素方差分析的主要部分。由于指定建立饱和模型,因此总的离差平方和分为3个部分:多个控制变量对观察量的独立作用、多个控制变量的交互作用及随机变量的影响。关于多个控制变量的独立作用部分,不同组别(教学方法)贡献离差平方和为3295.577,均方1647.788,不同性别贡献离差平方和为351.157,均方为351.157,这说明教学方法比性别影响大。从显著性概率来看,均小于0.05,说明两者均对数学成绩有影响。关于多个控制变量的交互作用分析类似,也对数学成绩具有显著性影响。误差部分是随机变量影响部分。

(4) “组别”变量的均值比较

表6.10是不同组别的均值比较结果(对比对话框中设置),可以看出不同组别之间的显著性概率值小于0.05,因此不同组别之间的均值具有显著性差异。

表 6.9 多因素方差分析及交互检验结果表 主体间效应的检验 因变量:数学

源	III 类平方和	自由度	均方	F	显著性
修正模型	4605.917a	5	921.183	17.163	.000
截距	95235.260	1	95235.260	1774.340	.000
组别	3295.577	2	1647.788	30.700	.000
性别	351.157	1	351.157	6.542	.025
组别 * 性别	599.843	2	299.922	5.588	.019
误差	644.083	12	53.674		
总计	112898.000	18			
修正后总计	5250.000	17			

a. R 方 = .877 (调整后 R 方 = .826)

表 6.10 “组别”变量的均值比较 对比结果 (K 矩阵)

组别 简单对比 a			因变量
			数学
级别 1 与级别 3	对比估算		16.625
	假设值		0
	差值 (估算 - 假设)		16.625
	标准误差		4.486
	显著性		.003
	差值的 95% 置信区间	下限	6.850
		上限	26.400
级别 2 与级别 3	对比估算		-17.500
	假设值		0
	差值 (估算 - 假设)		-17.500
	标准误差		4.360
	显著性		.002
	差值的 95% 置信区间	下限	-27.000
		上限	-8.000

a. 参考类别 = 3

(5) “性别”变量的均值比较

表 6.11 是对不同性别的均值比较结果, 由于显著性值不大于 0.05, 说明不同性别之间的均值有显著性差异。

表 6.11 “性别”变量的均值比较 对比结果 (K 矩阵)

性别 简单对比 a			因变量
			数学
级别 1 与级别 2	对比估算		-9.194
	假设值		0
	差值 (估算 - 假设)		-9.194
	标准误差		3.595
	显著性		.025
	差值的 95% 置信区间	下限	-17.026
		上限	-1.362

a. 参考类别 = 2

(6) “组别”变量的多重比较结果

表 6.12 是对组别进行多重比较的结果，由于前面分析方差具有齐性，从 LSD 结果可以看出 3 个水平的显著性均小于 0.05，说明三个组之间均存在显著性差异，表格中也用*标出了显著性差异，同时可看出其均值的比较性为第 0 组>第 2 组>第 1 组。

表 6.12 对组别变量的多重比较结果表 因变量：数学 LSD

(I) 组别	(J) 组别	平均值差值 (I-J)	标准误差	显著性	95% 置信区间	
					下限	上限
0	1	34.6667*	4.22980	.000	25.4507	43.8826
	2	12.8333*	4.22980	.010	3.6174	22.0493
1	0	-34.6667*	4.22980	.000	-43.8826	-25.4507
	2	-21.8333*	4.22980	.000	-31.0493	-12.6174
2	0	-12.8333*	4.22980	.010	-22.0493	-3.6174
	1	21.8333*	4.22980	.000	12.6174	31.0493

基于实测平均值。
误差项是均方（误差）= 53.674。
*. 平均值差值的显著性水平为 .05。

(7) 交互影响折线图

图 6-12 是两控制变量对观测变量的交互作用图，由于两因素相交，说明有交互作用的影响（这与对表 6.9 的分析结果一致）。

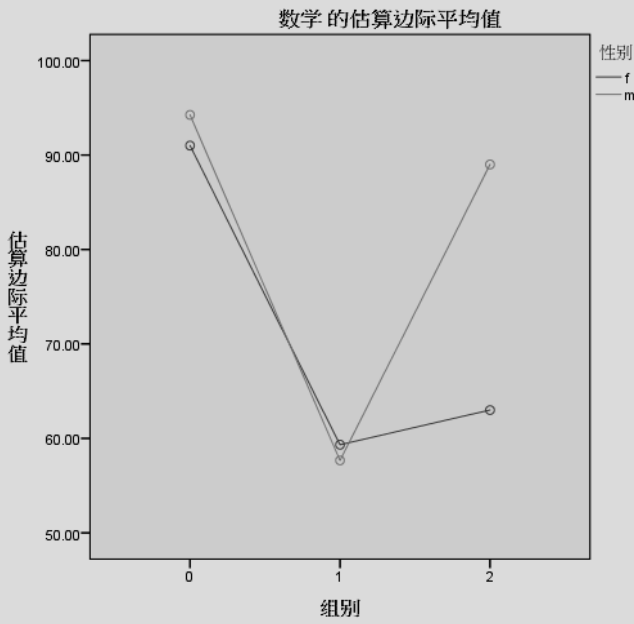


图 6-12 交互影响折线图

☆说明☆

◆ 当只有一个控制变量时，执行单变量过程等价于进行单因素方差分析。

6.4 协方差分析

6.4.1 基本概念及统计原理

1. 基本概念

无论是单因素方差分析还是多因素方差分析，都有一些可以人为控制的变量。在实际问题中，有些随机因素是很难人为控制的，但它们又会对结果产生显著的影响。如果忽略这些因素的影响，则有可能得不到正确的结论。

例如，研究某种药物对病症的治疗效果，如果仅仅分析药物本身的作用，而不考虑患者的身体素质，那么很可能得不到结论或得到的结论是错误的。

再如前面的例子，研究三种不同教学方法教学效果的好坏，检查教学效果是通过学生的考试成绩来体现的，而学生现在的考试成绩受他们自身知识基础的影响，在分析时就应排除这种影响。

为了更加准确地研究控制变量不同水平对结果的影响，应该尽量排除其他因素对分析结果的影响。例如，在上面的例子中，应尽量排除患者体质、学生数学基础的好坏等影响。为此就用到协方差分析。

协方差分析是将很难控制的因素作为协变量，在排除协变量影响的条件下，分析控制变量对观测变量的影响，从而更加准确地对控制因素进行分析和评价。协方差分析仍然沿袭方差分析的思想，并在分析观测变量离差时，考虑了协变量的影响，认为观测变量的变动受 4 个方面的影响，即控制变量的独立作用、控制变量的交互作用、协变量的作用和随机因素的作用，并在剔除协变量的作用后再分析控制变量对观测变量的影响。

☆说明☆

◆ 协方差分析要求协变量是连续数值型，多个协变量间相互独立，且与控制变量间无交互影响。

2. 统计原理

以单因素协方差分析为例，总的离差平方和表示为

$$Q_{\text{总}} = Q_{\text{控制变量}} + Q_{\text{协变量}} + Q_{\text{随机因素}} \tag{6.14}$$

协方差仍采用 F 检验， F 统计量的计算公式为

$$F_{\text{控制变量}} = \frac{S_{\text{控制变量}}^2}{S_{\text{随机因素}}^2} \tag{6.15}$$

$$F_{\text{协变量}} = \frac{S_{\text{协变量}}^2}{S_{\text{随机因素}}^2} \tag{6.16}$$

式中， S^2 表示相应变量的均方。

显而易见，如果相对于随机因素引起的离差，协变量带来的离差比较大，即 $F_{\text{协变量}}$ 值较大，则说明协变量是引起观测变量变动的主要因素之一，观测变量的变动可以部分地由协变量来线性解释；反之，则说明协变量没有给观测变量带来显著的线性影响。在排除了协变量的线性影响后，控制变量对观测变量的影响分析同方差分析。

3. 分析步骤

协方差分析问题也属于统计推断中的假设检验问题，其基本步骤与假设检验一致。

第1步 提出原假设。

协方差分析的原假设 H_0 ：协变量对观测变量的线性影响是不显著的；在扣除协变量影响的条件下，控制变量各水平下观测变量的总体均值无显著性差异，控制变量各水平对观测变量的效应同时为零。也就是说控制变量和协变量对观测变量均无显著性影响。

第2步 选择检验统计量。

协方差分析采用的是 F 统计量，计算公式如式（6.15）和式（6.16）所示。

第3步 计算检验统计量的观测值和概率 P 值。

SPSS 会根据式（6.15）和式（6.16）自动计算 F 统计值，并依据 F 分布表给出相应的显著性概率 P 值。协变量和控制变量对观测变量的显著性影响情况分析方法同前。

第4步 给出显著性水平 α ，作出决策。

如果显著性概率 P 值小于显著性水平 α ，则拒绝原假设，即认为控制变量不同水平下各总体均值有显著性差异；反之，认为控制变量不同水平下各总体均值没有显著性差异。

6.4.2 协方差分析 SPSS 实例分析

【例 6-3】 已知一个班三组学生的入学成绩和分别接受了三种不同教学方法后的数学成绩，如表 6.13 所示，试研究这三组学生在接受了不同的教学方法后数学成绩是否有显著性差异。（参见数据文件：data6-3.sav。）

表 6.13 三组学生的数学成绩

姓 名	数学	入学成绩	组别	姓 名	数学	入学成绩	组别
张青华	99	98	0	蔡春江	67	98	1
王洁云	88	89	0	武佳琪	56	78	1
吴凌风	99	80	0	陈雪吟	56	89	1
刘行	89	78	0	罗超波	79	87	2
马萌	94	78	0	尹珣	56	76	2
单玲玲	90	89	0	张敏	89	56	2
宋丽君	55	99	1	郭晓艳	99	76	2
辛瑞晶	50	89	1	李福利	70	89	2
王滢滢	67	88	1	罗帆	89	89	2

第1步 分析。

入学成绩肯定会对最后成绩有所影响，这里着重分析不同教学方法的影响，应将入学成绩（数学基础）的影响剔除，考虑用协方差分析。

第2步 数据组织。

将姓名、数学、入学成绩和组别分别定义为“name”、“math”、“entrance”和“group”，并设置其标签为中文名称，将数据输入并保存为文件 data6-3.sav。

第3步 检验协方差分析的前提条件。

与多因素方差分析操作一样，选择菜单“分析→一般线性模型→单变量”。该前提条件是各组方差具有齐性，以及协变量“入学成绩[entrance]”与控制变量“分组[group]”没有交互作用。因此将“数学成绩[math]”移入“因变量”框作为观测变量，将“分组[group]”移入“固定因子”框作为控制变量，将“入学成绩[entrance]”移入“协变量”框作为协变量，设置如图 6-13

所示。打开“模型”对话框，选中“定制”单选框自定义方差分析模型，并将“entrance”，“group”和“entrance*group”移入模型中（entrance*group 的移入方法是先同时选中“entrance”和“group”两个变量，再选择“构建项”中“类型”下拉列表框中的“交互”，就可通过向右的箭头移入“模型”框中），如图 6-14 所示。



图 6-13 协方差分析设置

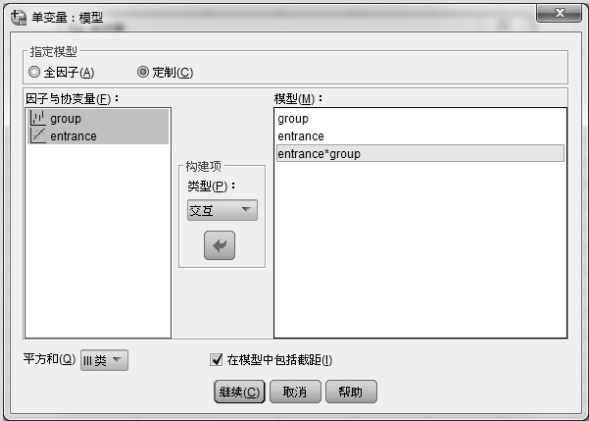


图 6-14 交互影响设置

打开如图 6-7 的“单变量：选项”对话框，选中“齐性检验”复选框，进行方差齐性检验。单击“确定”按钮运行，结果如表 6.14 和表 6.15 所示。

表 6.14 方差齐性检验结果 误差方差的莱文等同性检验^a 因变量:数学成绩

F	自由度 1	自由度 2	显著性
2.337	2	15	.131

检验“各个组中的因变量误差方差相等”这一原假设。

a. 设计: 截距 + group + entrance + group * entrance

表 6.14 是方差齐性检验结果。由于其显著性概率值 $P=0.131>0.05$ ，因此认为各组的方差具有齐性。

表 6.15 协变量与因变量交互作用检验 主体间效应的检验 因变量:数学成绩

源	III 类平方和	自由度	均方	F	显著性
修正模型	3757.122a	5	751.424	6.040	.005
截距	862.817	1	862.817	6.935	.022
group	104.163	2	52.082	.419	.667
entrance	.467	1	.467	.004	.952
group * entrance	61.932	2	30.966	.249	.784
误差	1492.878	12	124.406		
总计	112898.000	18			
修正后总计	5250.000	17			

a. R 方 = .716 (调整后 R 方 = .597)

表 6.15 主要是检验控制变量与协变量是否具有交互作用,从其中可看出 group 与 entrance 的交互作的显著性概率值 $P = 0.784 > 0.05$, 因此认为它们之间没有交互作用。

从以上分析可知,例 6-3 是满足协方差分析中关于方差齐性和协变量与控制变量之间没有交互作用这两个基本条件的,因此可用协方差分析来处理。

第 4 步 执行协方差分析。

变量的选择同第 3 步,只是在“模型”对话框之“定制”中的模型中不再选择 group*entrance。

第 5 步 主要结果及分析。

经过以上几个步骤的分析,得到协方差的主要结果,如表 6.16 所示。

表 6.16 协方差分析的主要结果 主体间效应检验 因变量:数学成绩

源	III 类平方和	自由度	均方	F	显著性
修正模型	3695.190a	3	1231.730	11.091	.001
截距	1387.824	1	1387.824	12.496	.003
group	3364.083	2	1682.041	15.146	.000
entrance	8.857	1	8.857	.080	.782
误差	1554.810	14	111.058		
总计	112898.000	18			
修正后总计	5250.000	17			

a. R 方 = .704 (调整后 R 方 = .640)

从表 6.16 可看出, group 所对应的显著性概率值 $P = 0.000 < 0.05$, 说明分组情况(不同的教学方法)对数学成绩具有显著性影响,而 entrance 所对应显著性概率值 $P = 0.782 > 0.05$, 说明入学成绩对最后的成绩无显著性影响。

☆说明☆

- ◆ 在做协方差分析前,需要对模型满足协方差分析的基本条件做出检验。同时,还需注意一点,在正式进行协方差分析时,一定不要将协变量和控制变量的交互作用纳入分析模型,否则可能产生完全相反的结论,因为协变量的作用是被排除在外的,不能与控制变量产生交互作用。

6.5 多元方差分析

6.5.1 基本概念及统计原理

1. 基本概念

多元方差分析是研究多个控制因素（自变量）与多个因变量相互关系的一种统计分析方法，又称为多变量分析。多元分析实质上是单变量统计方法的发展和推广，适用于研究控制因素同时对两个或两个以上的因变量产生影响的情况，用来分析控制因素取不同水平时这些因变量的均值是否存在显著性差异。

2. 统计原理

多元方差分析的基本原理同一元方差分析相似，是将总变异按照其来源（或实验设计）分为多个部分，从而检验各个因素对因变量的影响以及各因素间的交互作用。在这个过程中可以分析每个因素的作用，也可以分析因素之间的交互作用、协方差，以及各控制因素与协变量之间的交互作用。

多元方差分析的优点是可以在一次研究中同时检验具有多个水平的多个因素各自对因变量的影响以及各因素间的交互作用。

在方差分析中，要求样本必须满足独立、正态、等方差的总体，而对于多元方差分析而言，由于涉及多个因变量，除要求每个因变量满足以上条件外，还必须满足以下条件。

- ① 各因变量间具有相关性；
- ② 每一组都有相同的方差——协方差矩阵；
- ③ 各因变量为多元正态分布。

多元方差分析的目的在于检验控制因素如何影响一组因变量。SPSS 中用于多元方差分析假设检验的统计量有比莱（Pillai）轨迹、威尔克（Wilks）Lambda 值、霍特林（Hotelling）轨迹和罗伊（Roy）最大根。

- ① 比莱（Pillai）轨迹：该检验的值恒为正值，值越大表明该效应对模型的贡献越大。
- ② 威尔克（Wilks）Lambda 值：该检验的取值范围为 0~1，值越小表明该效应对模型的贡献越大。
- ③ 霍特林（Hotelling）轨迹：该值用于检验矩阵特征根之和，值越大表明该效应对模型的贡献越大。
- ④ 罗伊（Roy）最大根：该值用于检验矩阵特征根中的最大值，值越大表明该效应对模型的贡献越大。

如果方差齐性检验结果为显著性概率 P 值大于 0.05，则方差齐性假设成立，就用威尔克（Wilks）Lambda 的检验值进行判断；反之，则用其他几项检验数据进行判断。

3. 分析步骤

多元方差分析的步骤与单因素方差分析和协方差分析比较相近，下面通过具体实例来说明。

6.5.2 多元方差分析 SPSS 实例分析

【例 6-4】某科研所研究某树种在不同海拔、不同施肥量情况下的苗高增加量和地径增加量的差别，将海拔设为 3 个水平，并将施肥量也设为 3 个水平，将两个因素组合成 9 个组合，每

个组合重复 3 次。试分析海拔和施肥量对苗高增加量和地径增加量的影响，并分析海拔与施肥量是否存在交互作用。（数据来源：李昕等，《SPSS 20.0 统计分析从入门到精通》，电子工业出版社；参见数据文件：data6-4.sav。）

表 6.17 某树种的生长数据

海拔	施肥量	苗高增加量	地径增加量	海拔	施肥量	苗高增加量	地径增加量
1	1	11.8	2.48	2	2	9.9	1.88
1	1	12.9	2.7	2	3	8.7	1.59
1	1	10.9	2.84	2	3	9	1.93
1	2	9.6	2.07	2	3	8.3	1.84
1	2	9.4	1.88	3	1	11.1	1.9
1	2	9.1	1.72	3	1	10.8	1.8
1	3	8.2	1.19	3	1	10.2	2.1
1	3	8.8	1.79	3	2	9.1	1.6
1	3	9.1	1.9	3	2	8.8	1.68
2	1	11.3	2.12	3	2	8.5	1.9
2	1	10.6	1.97	3	3	7.3	1.4
2	1	11.7	2.53	3	3	8.1	1.6
2	2	10.1	2.01	3	3	8.2	1.79
2	2	10.4	1.93				

第 1 步 分析。

这是一个两个控制因素对两个因变量影响的分析，是一个多元方差分析问题。

第 2 步 数据组织。

按表 6.17 的变量名组织成 4 列数据，并将数据文件保存为 data6-4.sav。

第 3 步 分析过程设置。

选择菜单“分析→一般线性模型→多变量”。将“苗高增加量”和“地径增加量”移入因变量框，将“海拔”和“施肥量”移入固定因子框。打开“事后比较 (H) ...”对话框，将“海拔”和“施肥量”移入到“下列各项的事后检验”列表框，并勾选“假定等方差”选项组中的“LSD”复选框。打开“选项 (O) ...”对话框，在“显示”选项组中，勾选“齐性检验”复选框。完成设置并运行。

第 4 步 主要结果及分析。

运行结果如表 6.18~表 6.22 所示，对各表的具体分析如下。

(1) 因变量的方差齐性检验结果

从表 6.18 可看出，苗高增加量和地径增加量的显著性概率 P 值分别为 0.344 和 0.166，均大于显著性水平 0.05，说明两者在各组总体方差具有齐性，满足方差分析的前提条件。

表 6.18 误差方差的莱文等同性检验^a

	F	自由度 1	自由度 2	显著性
苗高增加量	1.216	8	18	.344
地径增加量	1.703	8	18	.166

检验“各个组中的因变量误差方差相等”这一原假设。

a. 设计：截距 + 海拔 + 施肥量 + 海拔 * 施肥量

(2) 多元方差分析结果

表 6.19 是多元变量检验结果, 可看出海拔与施肥量两个主效应的 4 种检验显著性概率均小于 0.05, 说明海拔与施肥量对苗高增加量和地径增加量有显著性影响; 而“海拔*施肥量”的 4 种检验的显著性概率均大于 0.05, 说明两者对苗高增加量和地径增加量的影响不存在交互作用。

表 6.19 多变量检验^a 结果

效应		值	F	假设自由度	误差自由度	显著性
截距	比莱轨迹	.998	4789.516b	2.000	17.000	.000
	威尔克 Lambda	.002	4789.516b	2.000	17.000	.000
	霍特林轨迹	563.472	4789.516b	2.000	17.000	.000
	罗伊最大根	563.472	4789.516b	2.000	17.000	.000
海拔	比莱轨迹	.580	3.673	4.000	36.000	.013
	威尔克 Lambda	.443	4.278b	4.000	34.000	.007
	霍特林轨迹	1.210	4.839	4.000	32.000	.004
	罗伊最大根	1.167	10.503c	2.000	18.000	.001
施肥量	比莱轨迹	.902	7.395	4.000	36.000	.000
	威尔克 Lambda	.106	17.666b	4.000	34.000	.000
	霍特林轨迹	8.404	33.616	4.000	32.000	.000
	罗伊最大根	8.396	75.560c	2.000	18.000	.000
海拔 * 施肥量	比莱轨迹	.586	1.864	8.000	36.000	.097
	威尔克 Lambda	.491	1.814b	8.000	34.000	.109
	霍特林轨迹	.880	1.759	8.000	32.000	.123
	罗伊最大根	.632	2.843c	4.000	18.000	.055

a. 设计: 截距 + 海拔 + 施肥量 + 海拔 * 施肥量

b. 精确统计

c. 此统计是生成显著性水平下限的 F 的上限。

(3) 主体间效应的检验结果

表 6.20 是两个因变量在不同影响因素上的差异分析。可看出, 苗高增加量在海拔和施肥量上的显著性概率分别为 0.002 和 0.000, 说明苗高增加量在海拔和施肥量上均存在显著性差异; 地径增加量在海拔和施肥量上的显著性概率分别为 0.018 和 0.000, 说明地径增加量在海拔和施肥量上均存在显著性差异; 而苗高增加量与地径增加量在“海拔*施肥量”上的显著性概率为 0.237 和 0.058, 均大于 0.05, 说明海拔与施肥量的交互作用在苗高增加量与地径增加量上均无显著性差异。这与表 6.19 的分析情况相吻合。

表 6.20 主体间效应的检验结果

源	因变量	III 类平方和	自由度	均方	F	显著性
修正模型	苗高增加量	43.447a	8	5.431	21.098	.000
	地径增加量	2.687b	8	.336	7.293	.000
截距	苗高增加量	2540.430	1	2540.430	9869.296	.000
	地径增加量	100.688	1	100.688	2186.056	.000
海拔	苗高增加量	4.509	2	2.254	8.758	.002
	地径增加量	.465	2	.232	5.047	.018

续表

源	因变量	III 类平方和	自由度	均方	F	显著性
施肥量	苗高增加量	37.369	2	18.684	72.587	.000
	地径增加量	1.710	2	.855	18.563	.000
海拔 * 施肥量	苗高增加量	1.569	4	.392	1.524	.237
	地径增加量	.512	4	.128	2.781	.058
误差	苗高增加量	4.633	18	.257		
	地径增加量	.829	18	.046		
总计	苗高增加量	2588.510	27			
	地径增加量	104.205	27			
修正后总计	苗高增加量	48.080	26			
	地径增加量	3.516	26			

a. R 方 = .904 (调整后 R 方 = .861)

b. R 方 = .764 (调整后 R 方 = .659)

(4) 多重比较结果分析

表 6.21 是海拔的多重比较结果,可看出苗高增加量在海拔 1 与 2、1 与 3、2 与 3 上的显著性概率分别为 0.927、0.002 和 0.002,说明苗高增加量在海拔 1 与 3、2 与 3 上存在显著性差异,在 1 与 2 上没有显著性差异;同时,可看出地径增加量在海拔 1 与 3、2 与 3 上存在显著性差异,而在 1 与 2 上没有显著性差异。

表 6.21 海拔的多重比较结果

LSD

因变量	(I) 海拔	(J) 海拔	平均值差值 (I-J)	标准误差	显著性	95% 置信区间	
						下限	上限
苗高增加量	1	2	-.022	.2392	.927	-.525	.480
		3	.856*	.2392	.002	.353	1.358
	2	1	.022	.2392	.927	-.480	.525
		3	.878*	.2392	.002	.375	1.380
	3	1	-.856*	.2392	.002	-1.358	-.353
		2	-.878*	.2392	.002	-1.380	-.375
地径增加量	1	2	.0856	.10117	.409	-.1270	.2981
		3	.3111*	.10117	.007	.0986	.5237
	2	1	-.0856	.10117	.409	-.2981	.1270
		3	.2256*	.10117	.039	.0130	.4381
	3	1	-.3111*	.10117	.007	-.5237	-.0986
		2	-.2256*	.10117	.039	-.4381	-.0130

基于实测平均值。

误差项是均方 (误差) = .046。

*. 平均值差值的显著性水平为 .05。

表 6.22 是施肥量的多重比较结果,可看出苗高增加量在施肥量 1 与 2、1 与 3 和 2 与 3 上均存在显著性差异;地径增加量在施肥量 1 与 2、1 与 3 上存在显著性差异,而在 2 与 3 上没有显著性差异。

表 6.22 施肥量的多重比较结果

LSD								
因变量	(I) 施肥量	(J) 施肥量	平均值差值 (I-J)	标准误差	显著性	95% 置信区间		
						下限	上限	
苗高增加量	1	2	1.822*	.2392	.000	1.320	2.325	
		3	2.844*	.2392	.000	2.342	3.347	
	2	1	-1.822*	.2392	.000	-2.325	-1.320	
		3	1.022*	.2392	.000	.520	1.525	
	3	1	-2.844*	.2392	.000	-3.347	-2.342	
		2	-1.022*	.2392	.000	-1.525	-.520	
地径增加量	1	2	.4189*	.10117	.001	.2063	.6314	
		3	.6011*	.10117	.000	.3886	.8137	
	2	1	-.4189*	.10117	.001	-.6314	-.2063	
		3	.1822	.10117	.088	-.0303	.3948	
	3	1	-.6011*	.10117	.000	-.8137	-.3886	
		2	-.1822	.10117	.088	-.3948	.0303	

基于实测平均值。
误差项是均方（误差）=.046。
*. 平均值差值的显著性水平为 .05。

6.6 典型案例

6.6.1 培训材料效果分析

为了研究三种不同培训材料对强化员工全面质量管理意识的作用是否有显著差异，从某企业随机选择了 18 名员工，并将他们随机划分为 3 组，每组分别采用不同的培训材料进行培训。培训结束后对他们进行考试，其所得的考试分数如表 6.23 所示，试分析使用不同培训材料的培训效果是否存在显著性差异。（参见数据文件：data6-5.sav。）

案例分析：要分析不同培训材料的培训效果是否存在显著差异，就要检验使用不同培训材料后考试成绩的均值是否有显著差异。本例中只有一个因素（培训材料），即一个自变量，它有 3 个水平（即自变量有 3 个值），分别是材料 1、2、3。因变量为考试分数。分析前，先进行方差齐性检验，然后再使用单因素方差分析方法来判断。

表 6.23 使用不同培训材料的培训效果

材料 1	材料 2	材料 3
85	71	59
75	75	64
82	73	62
76	74	69
71	69	75
85	82	67

6.6.2 火箭射程影响因素分析

为了研究火箭燃料和推进器对火箭射程的影响，选用了 4 种不同燃料和 3 种型号的火箭推进器，将它们相互搭配并在每一种搭配情况下做了两次试验，得到火箭射程（海里）数据，如表 6.24 所示。试分析燃料和推进器这两种因素对火箭射程的影响是否显著。（数据来源：郝黎仁等，《SPSS 实用统计分析》，中国水利水电出版社；参见数据文件：data6-6.sav。）

案例分析：燃料和推进器是影响火箭射程的重要因素（双因素），但究竟哪种燃料、哪种推进

器或者哪种燃料和推进器的组合影响最显著，就需用双因素方差分析来解决。通过主效应效果检查表可以看出主效应及其交互效应是否显著，通过两两对比表可以看出具体哪种燃料和哪种推进器效果显著，然后通过轮廓图查看交互效应情况。

表 6.24 火箭射程（海里）数据

推进器 B 燃料 A			
	B1	B2	B3
A1	58.20, 52.60	56.20, 41.20	65.30, 60.80
A2	49.10, 42.80	54.10, 50.50	51.60, 48.40
A3	60.10, 58.30	70.90, 73.20	39.20, 40.70
A4	75.80, 71.50	58.20, 51.00	48.70, 41.40

6.6.3 三种治疗高血压病的方法效果分析

某高血压研究中心开发了三种治疗高血压的方法，表 6.25 所示为患者采用不同治疗方法后的血压，试分析这三种方法是否有显著差异。（数据来源：宋志刚等，《SPSS 16 实用教程》，人民邮电出版社；参见数据文件：data6-7.sav。）

表 6.25 三组病人的血压

患者编号	治疗后的血压	治疗前的血压	组别	患者编号	治疗后的血压	治疗前的血压	组别
1	120	160	0	10	110	150	1
2	125	185	0	11	125	155	1
3	130	155	0	12	125	155	1
4	150	145	0	13	105	160	2
5	145	175	0	14	150	175	2
6	160	175	0	15	145	165	2
7	135	180	1	16	140	155	2
8	140	210	1	17	125	190	2
9	125	220	1	18	110	165	2

案例分析：患者入院前的血压对其治疗效果肯定有影响，但要分析三种治疗方法的效果，必须先把入院前的血压对其治疗效果的影响剔除，因此考虑用协方差分析法。先要判断各组方差是否齐性，再判断协变量（入院前的血压）与因素（治疗方法）之间有没有交互作用。满足这两个条件后才可以协方差分析的方法来处理。

6.7 思考与练习

1. 方差分析是用来检验不同数据组均值差异的，还是检验方差差异的？
2. 如果单因素方差分析的结果是：不同方案的效果均值有显著性差异，是否意味着两两方案之间的均值都有显著性差异？
3. 方差分析假定的前提条件有哪些？什么是主效应？什么是交互效应？
4. 为了寻求适应某地区的高产油菜品种，今选了 5 种品种进行试验，每一品种在 4 块条件完全相同的试验田上试种，其他施肥等田间管理措施完全一样。表 6.26 所示为每一品种下每一块田的亩产量，根据这些数据分析不同品种油菜的平均产量在显著性水平 0.05 下有无显著性差异。（数据来源：王玲玲，《常用统计方法》，华东师范大学出版社；参见数据文件：data6-8.sav。）
5. 某公司希望检测四种类型轮胎 A、B、C、D 的寿命（由行驶的里程数决定），如表 6.27 所示（单

位：千英里），其中每种轮胎应用在随机选择的 6 辆汽车上。在显著性水平 0.05 下判断不同类型轮胎的寿命间是否存在显著性差异？（数据来源：M.R.斯皮格尔，《统计学（第三版）》，科学出版社；参见数据文件：data6-9.sav。）

表 6.26 小麦产量的实测数据

品种	A1	A2	A3	A4	A5
亩产量	256.0	244.0	250.0	288.0	206.0
	222.0	300.0	277.0	280.0	212.0
	280.0	290.0	230.0	315.0	220.0
	298.0	275.0	322.0	259.0	212.0

表 6.27 四种轮胎的寿命数据

A	B	C	D
33	32	31	29
38	40	37	34
36	42	35	32
40	38	33	30
31	30	34	33
35	34	30	31

6. 某超市将同一种商品做 3 种不同的包装（A）并摆放在 3 个不同的货架区（B）进行销售试验，随机抽取 3 天的销售量作为样本，具体资料见表 6.28。要求检验：在显著性水平 0.05 下商品包装、摆放位置及其搭配对销售情况是否有显著性影响。（数据来源：耿修林，《应用统计学》，科学出版社；参见数据文件：data6-10.sav。）

表 6.28 销售样本数据

	B1	B2	B3
A1	5,6,4	6,8,7	4,3,5
A2	7,8,8	5,5,6	3,6,4
A3	3,2,4	6,6,5	8,9,6

7. 研究杨树一年生长量与施用氮肥和钾肥的关系。为了研究这种关系，共进行了 18 个样地的栽培试验，测定杨树苗的一年生长量、初始高度、全部试验条件（包括氮肥量和钾肥量）及试验结果（杨树苗的生长量）数据，如表 6.29 所示，请在显著性水平 0.05 下检验氮肥量、钾肥量及树苗初始高度中哪些对杨树的生长有显著性影响。（数据来源：李勇，《生物数学模型的统计学基础》，科学出版社；参见数据文件：data6-11.sav。）

表 6.29 杨树栽培试验数据

序号	氮肥量	钾肥量	树苗初高	生长量	序号	氮肥量	钾肥量	树苗初高	生长量
1	少	0	4.5	1.85	10	多	0	6.5	2.15
2	少	0	6	2	11	多	0	6	1.99
3	少	0	4	1.6	12	多	0	6.5	2.06
4	少	12.5	6.5	2	13	多	12.5	4	1.93
5	少	12.5	7	2.04	14	多	12.5	6	2.1
6	少	12.5	5	1.91	15	多	12.5	5.5	2.15
7	少	25	7	2.4	16	多	25	5	2.2
8	少	25	5	2.25	17	多	25	6	2.3
9	少	25	5	2.1	18	多	25	5.5	2.25

第7章 相关分析

在前面几章中，讲解的方法基本上都是一元统计方法，从本章开始，介绍多元统计分析的模型和方法。多元统计分析方法是分析多个性质不同的 SPSS 变量，分析总体的多个特征，并分析这些特征的联系。相关分析是比较简单的多元分析方法，但也是经常使用的多元统计分析方法，能快速发现总体特征之间的关系，并检验这些特征的显著性。近年来，相关分析广泛应用于生物学、心理学、教育学、经济学、医学等各个领域。相关分析对于实验数据的处理、经验公式的建立、管理标准的测定、自然现象和经济现象的统计预报、自动控制中数学模型的确定等，是一种极为有效且广泛使用的数理统计工具。

本章通过例子学习相关分析及其在 SPSS 中的实现。

7.1 相关分析简介

7.1.1 相关分析的概念

客观世界是普遍联系的统一整体，事物之间存在相互依赖、相互制约、相互影响的关系。描述事物数量特征的变量之间自然也存在一定的关系，变量之间的关系可以分为两种：一种是函数关系，另一种是相关关系。

函数关系是一种一一对应的关系，即当一个变量 x 取一定值时，另一变量 y 可以按照确定的函数取一个确定的值，记为 $y=f(x)$ ，则称 y 是 x 的函数，也就是说， y 与 x 两变量之间存在函数关系。例如，在单价确定的条件下，给定销售量就能确定销售额；圆的周长和圆的半径的关系等。

函数关系是一一对应的确定性关系，比较容易分析和测度。可是在现实世界中，变量间的关系往往并不是简单的确定性关系，也就是说，变量之间有着密切的关系，但又不能由一个或几个变量的值确定另一个变量的值，即当自变量 x 取某一值时，因变量 y 的值可能会有多个。这种变量之间非一一对应的、不确定性的关系，称为相关关系。例如，子女身高与父母身高之间的关系，虽然两者之间存在一定的关系，但这种关系却不能像函数关系那样用一个确定的数学函数来描述。

7.1.2 相关关系的种类

1. 按相关关系涉及的变量数量分类

相关关系按照涉及的变量个数，可以分为简单相关和复相关两种。简单相关是指一个变量和另一个变量之间的相关关系，例如，人的身高与体重之间的相关关系。复相关是指一个变量和另一组变量之间的相关关系，例如，某种商品的需求量与商品的价格及居民的收入水平之间的相关关系。

2. 按变量相关关系的表现形式分类

相关关系按照表现形式的不同，分为线性相关和非线性相关两种。线性相关是指一个变量变

化时，其变化量与另一个变量的变化量有大致按比例的变化，两个变量的散点图近似落在一条直线附近。当变量之间相关关系散点图中的点接近于一条曲线时，称为非线性相关，又称为曲线相关。

3. 按变量相关关系变化的方向分类

相关关系按照相关方向的不同，分为正相关和负相关两种。当两个变量趋于在同一个方向变化时，即同增或同减，称为变量之间存在正相关。当两个变量趋于在相反方向变化时，即当一个变量增加时，另一个变量却减少，称为变量之间存在负相关。

4. 按变量相关的程度分类

相关关系按照相关程度，分为不相关、低度相关、显著相关、高度相关和完全相关。当一个变量的变化完全由另一个变量的变化所确定时，称为变量之间完全相关。例如，在价格不变条件下，某种商品销售额与销售量之间的关系，在这种情况下，相关关系实际成为了函数关系，所以可以把函数关系视为相关关系的特例。

当两个变量的变化之间完全没有关系，即彼此互不影响时，称为二者不相关。低度相关、显著相关和高度相关介于完全相关和不相关之间时，统称为不完全相关。

7.2 两变量相关分析

7.2.1 基本概念及统计原理

1. 相关系数

在各种相关分析中，只有两个变量的线性相关关系的分析是最简单的。两个变量之间的线性相关程度可以用简单线性相关系数去度量，相关系数是反映变量之间相关关系密切程度的统计量，根据线性相关系数计算方法的不同，线性相关系数具体分为如下三种。

(1) 皮尔逊(Pearson) 相关系数

这是最简单也最常用的相关系数，用于衡量间隔尺度变量间的线性关系。其计算公式如下：

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

(7.1)

式(7.1)只是代表了样本的相关系数，其中， n 为样本数， x_i, y_i 代表两个变量的样本观测值，它有如下特点：

- 相关系数 r 的取值在 $-1 \sim 1$ 之间，当 $0 < |r| < 1$ 时，表明 X 和 Y 之间存在一定的线性相关关系。若 $r > 0$ ，表明 X 和 Y 完全正相关；若 $r < 0$ ，表明 X 和 Y 负相关。
- 当 $r = 0$ 时，表明 X 和 Y 没有线性相关关系。
- 当 $|r| = 1$ 时，表明 X 和 Y 完全线性相关。若 $r = 1$ ，表明 X 和 Y 完全正相关；若 $r = -1$ ，表明 X 和 Y 完全负相关。
- x, y 对称， x, y 变量互换位置， r 不变，即 $r_{XY} = r_{YX}$ 。
- r 是标准化后计算的，因此是无量纲数。

☆说明☆

- (1) 皮尔逊 (Pearson) 相关系数适用于两变量的度量水平都是间隔尺度数据, 两变量的总体是正态分布或近似分布的情况, 否则其反映的线性关系有可能失真;
- (2) 相关系数为 0 或接近于 0 时, 只能说明没有线性关系, 不能说两个变量之间没有相关性, 有可能存在其他非线性关系。

(2) 斯皮尔曼 (Spearman) 相关系数

在进行相关分析的过程中, 我们经常会遇到一些不适宜用皮尔逊相关系数的数据, 例如, 变量的度量尺度不是间隔尺度而是顺序尺度的数据, 变量总体的分布不详, 这时用皮尔逊相关系数就不再适用。

若两列变量值为顺序尺度的数据 (又称为定序数据), 并且变量值所属的两个总体并不一定呈正态分布, 样本容量不一定大于 30, 这时两个变量之间的相关性可以通过计算斯皮尔曼相关系数进行分析。斯皮尔曼相关系数的计算公式为

$$r = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)} \quad (7.2)$$

式中, n 为样本容量; $\sum_{i=1}^n D_i^2 = \sum_{i=1}^n (U_i - V_i)^2$, 这里的 (U_i, V_i) 是两变量的秩。

斯皮尔曼相关系数的适用条件:

- 两个变量为度量尺度是顺序尺度的变量。
- 样本容量 n 不一定大于 30, 两个变量的总体不一定呈正态分布。

(3) 肯德尔 tau-b (Kendall τ) 等级相关系数

肯德尔 tau-b 等级相关系数计算仍基于数据的秩, 利用变量的秩计算一致对数目 U 和非一致对数目 V 。例如, 两变量 (x_i, y_i) 的秩对分别为 (2, 3)、(4, 4)、(3, 1)、(5, 5)、(1, 2), 对变量 x 的秩按升序排列后的秩对为 (1, 2)、(2, 3)、(3, 1)、(4, 4)、(5, 5), 于是, 变量 y 的秩随变量 x 的秩同步增大的秩对 (一致对) 有 (2, 3)、(2, 4)、(2, 5)、(3, 4)、(3, 5)、(1, 4)、(1, 5)、(4, 5), 一致对数目 U 等于 8; 变量 y 的秩未随变量 x 的秩同步增大的秩对 (非一致对)

有 (2, 1)、(3, 1), 非一致对数目 V 等于 2。于是, 一致对数目定义为 $U = \sum_{i=1}^n \sum_{j>i}^n I(d_j > d_i)$, 非

一致对数目定义为 $V = \sum_{i=1}^n \sum_{j>i}^n I(d_j < d_i)$ 。显然, 当一致对数目较大、非一致对数目较小时, 两变

量呈较强的正相关; 当一致对数目较小、非一致对数目较大时, 两变量呈较强的负相关; 当一致对数目和非一致对数目接近时, 两变量呈较弱的相关关系。

肯德尔 tau-b 等级相关系数的计算公式为

$$\tau = (U - V) \frac{2}{n(n-1)} \quad (7.3)$$

2. 相关系数的显著性检验

样本相关系数是根据从总体中抽取的随机样本的观测值 x 和 y 计算出来的, 它只是对总体相关系数 ρ 的估计。由于不同的样本可以计算出同一个样本相关系数, 因此样本相关系数不是一个

确定的值，而是随抽样变动的随机变量。那么，我们所估计的样本相关系数是否为抽样的偶然结果呢？为此，相关系数的统计显著性还有待检验。

对相关系数的显著性检验通常是检验总体相关系数是否等于零，对于不同的相关系数，其统计检验的统计量也不相同，构建的假设检验也略有差异，下面分别介绍。

(1) 皮尔逊相关系数假设检验

检验的原假设是总体相关系数 $\rho = 0$ ，即相关系数不显著，在原假设为真的条件下，与样本相关系数 r 有关的 t 统计量服从自由度为 $(n - 2)$ 的 T 分布：

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{7.4}$$

SPSS 会自动计算 T 检验统计量的观测值和对应的显著性概率 P 值，根据 P 值来判断相关系数的显著性。

(2) 斯皮尔曼相关系数假设检验

检验的原假设也是总体相关系数 $\rho = 0$ ，在小样本下，斯皮尔曼相关系数 r 就是检验统计量，在大样本时，采用正态检验统计量 Z 统计量，即

$$Z = r\sqrt{n-1} \tag{7.5}$$

式中， Z 统计量服从标准正态分布。SPSS 将自动计算斯皮尔曼相关系数、 Z 检验统计量的观测值和对应的概率 P 值。

(3) 肯德尔 tau-b 等级相关系数假设检验

检验的原假设也是总体相关系数 $\rho = 0$ ，在小样本情况下，肯德尔 tau-b 等级相关系数 τ 就是检验统计量，在大样本情况下采用的检验统计量为

$$Z = \tau \sqrt{\frac{9n(n-1)}{2(2n+5)}} \tag{7.6}$$

式中， Z 统计量近似服从标准正态分布。SPSS 将自动计算肯德尔 tau-b 等级相关系数、 Z 检验统计量和对应的概率 P 值。

以上三种相关系数的显著性检验，都可以根据 SPSS 计算出的显著性概率 P 值和显著性水平比较来完成。

7.2.2 两变量相关分析 SPSS 实例分析

【例 7-1】 为了分析父亲与儿子身高之间的相关性，现抽样了 12 对父子的身高，数据如表 7.1 所示。请对其进行相关性分析（显著性水平取 $\alpha = 0.05$ ）。（数据来源：M.R.斯皮格尔，《统计学（第三版）》，科学出版社；参见数据文件：data7-1.sav。）

表 7.1 12 对父子的身高数据（单位：英寸）

父亲身高	65	63	67	64	68	62	70	66	68	67	69	71
儿子身高	68	66	68	65	69	66	68	65	71	67	68	70

第 1 步 分析。

由于考虑的是父亲和儿子身高的相关性问题，故应用二元变量的相关性进行分析，同时身高是定距变量，考虑用皮尔逊相关系数来衡量。

第 2 步 数据的组织。

数据分成两列，一列是父亲的身高，变量名为“father”；另一列是儿子的身高，变量名为

“son”，输入数据并保存。

第3步 两变量的相关性分析。

选择菜单“分析→相关→双变量”，打开如图 7-1 所示的对话框，将“father”和“son”两变量移入“变量”框中；“相关系数”选择“皮尔逊”；在“显著性检验”中选择“双尾”；单击“选项(O)…”按钮，弹出如图 7-2 所示的对话框，选中“统计”选项框下的两项，计算结果中将输出均值和标准差、叉积偏差和协方差。



图 7-1 “双变量相关性”对话框



图 7-2 “双变量相关性：选项”对话框

第4步 主要结果及分析。

运行的主要结果如表 7.2 和表 7.3 所示，具体分析如下。

(1) 描述性统计量

表 7.2 列出了描述性统计量平均值、标准差和统计量个案数。

表 7.2 描述性统计量

描述统计			
	平均值	标准差	个案数
父亲身高	66.67	2.774	12
儿子身高	67.58	1.881	12

(2) 相关分析结果表

表 7.3 是相关分析的主要结果，其中包括平方和与叉积、协方差、皮尔逊相关系数及显著性概率 P 值。从表中可看出，相关系数为 $0.703 > 0$ ，说明呈正相关，相关系数的显著性为 $0.011 < 0.05$ ，因此应拒绝原假设 (H_0 : 两变量之间相关系数为零)，即说明儿子身高受父亲身高显著性正影响。从表下的注释可看出，两变量在 0.05 水平上显著相关。

表 7.3 双变量相关性检验结果

相关性			
		父亲身高	儿子身高
父亲身高	皮尔逊相关性	1	.703*
	显著性（双尾）		.011
	平方和与叉积	84.667	40.333
	协方差	7.697	3.667
	个案数	12	12
儿子身高	皮尔逊相关性	.703*	1
	显著性（双尾）	.011	
	平方和与叉积	40.333	38.917
	协方差	3.667	3.538
	个案数	12	12
* 在 .05 级别（双尾），相关性显著。			

☆说明☆

- (1) 对于单尾检验和双尾检验的选择一般遵循的原则是：如果不清楚变量之间是正相关还是负相关，应选择双尾检验；如果了解变量之间是正相关或负相关，则应选择单尾检验。
- (2) 可以看出表 7.3 的相关系数矩阵是一个对称矩阵，父亲身高与儿子身高和儿子身高与父亲身高的相关系数是一样的，从这里可以看出：对相关性来讲，两变量之间的地位是平等的，无主次之分。

【例 7-2】 1990 年中国科协管理科学研究中心在中国公众对待科学技术态度的问卷调查中，列举了 12 种职业，要求被调查者对声望高低和值得信赖程度进行回答，据回收的答卷按照公众对各种职业态度的人数排序，取得数据如表 7.4 所示，根据这些数据计算等级相关系数，并检验其显著性。（数据来源：郝黎仁，《SPSS 实用统计分析》，中国水利水电出版社；参见数据文件：data7-2.sav。）

表 7.4 公众对待 12 种社会职业的评价态度数据表

职业	社会声望	值得信赖程度
科学家	1	1
医生	2	2
政府官员	3	7
工程师	6	4
大学教师	5	5
律师	8	6
记者	7	8
建筑设计人员	11	9
银行管理人员	10	10
会计师	12	11
企业管理人员	9	12
中小学教师	4	3

第 1 步 分析。

由于 12 种职业的社会声望及值得信赖的程度均是定序数据，故考虑用斯皮尔曼相关系数进行分析。

第 2 步 数据的组织。

数据分成三列，第一列是职业，变量名为“job”；第二列是社会声望，变量名为“renown”；第三列是值得信赖程度，变量名为“confide”，输入数据并保存。

第 3 步 两元变量的相关性分析。

选择菜单：“分析→相关→双变量”，打开如图 7-1 所示的对话框，将“renown”和“confide”两变量移入“变量”框中；相关系数选择“斯皮尔曼”和“肯德尔 tau-b”系数；在显著性检验中选择双尾检验。

第 4 步 主要结果及分析。

运行的主要结果如表 7.5 所示。

具体分析如下：

从表 7.5 的上半部分可看出，两变量的肯德尔相关系数为 $0.697 > 0$ ，双尾检验的显著性概率为 $0.002 < 0.05$ ，应拒绝两变量不相关的原假设，说明两变量具有显著的正相关性。

从表 7.5 的下半部分可看出，两变量的斯皮尔曼相关系数为 $0.860 > 0$ ，同时双尾检测的显著性概率值 $P=0.000<0.05$ ，也说明两变量呈显著的正相关性。从表的脚注可看出双尾检测下两变量在 0.01 水平上具有显著的正相关性。

表 7.5 双变量相关性检验结果

相关性				
			社会声望	值得信赖程度
肯德尔 tau_b	社会声望	相关系数	1.000	.697**
		显著性（双尾）	.	.002
		个案数	12	12
	值得信赖程度	相关系数	.697**	1.000
		显著性（双尾）	.002	.
		个案数	12	12
斯皮尔曼 Rho	社会声望	相关系数	1.000	.860**
		显著性（双尾）	.	.000
		个案数	12	12
	值得信赖程度	相关系数	.860**	1.000
		显著性（双尾）	.000	.
		个案数	12	12
**. 在 0.01 级别（双尾），相关性显著。				

7.3 偏相关分析

7.3.1 基本概念及统计原理

1. 基本概念

相关分析计算两个变量之间的相互关系，分析两个变量间线性相关的程度，往往因为第三个变量所起的作用，使得相关系数不能真实地反映两个变量间的线性相关程度，这就导致了二元变量相关分析的不精确性。例如身高、体重与肺活量之间的关系，如果用皮尔逊相关分析计算其相关系数，可以得出肺活量与身高、体重均存在较强的线性关系。但实际上，如果对体重相同的人，分析身高和肺活量，是否身高值越大，肺活量也越大呢？结论是否定的。正是因为身高与体重之间存在线性关系，体重与肺活量之间存在线性关系，而得出身高与肺活量之间存在线性关系的错误结论。

偏相关分析的任务就是在研究两个变量之间的线性相关关系时控制可能对其产生影响的变量，这种相关系数称为偏相关系数。偏相关系数的数值和简单相关系数的数值常常是不同的，在计算简单相关系数时，所有其他自变量不予考虑。在计算偏相关系数时，要考虑其他自变量对因变量的影响，只不过是把其他自变量当作常数处理了。

根据观测资料应用偏相关分析计算偏相关系数，可以判断哪些自变量对因变量的影响较大，而选择作为必须考虑的自变量。至于那些对因变量影响较小的自变量，则可舍去。这样在计算多元回归分析时，只需保留起主要作用的自变量，用较少的自变量描述因变量的平均变动量。偏相关分析在自然科学和社会科学的各个方面都有着非常广泛的应用。

2. 统计原理

控制变量为 z ，变量 x 、 y 之间的偏相关系数定义为

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \tag{7.7}$$

式中， $r_{xy,z}$ 是在控制 z 的条件下， x 、 y 之间的偏相关系数； r_{xy} 是变量 x 、 y 之间的简单相关系数， r_{xz} 是变量 x 、 z 之间的简单相关系数， r_{yz} 是变量 y 、 z 之间的简单相关系数。

当控制两个变量 z_1 、 z_2 时，变量 x 、 y 之间的偏相关系数计算公式为

$$r_{xy,z} = \frac{r_{xy,z_1} - r_{xz_1,z_2}r_{yz_2,z_1}}{\sqrt{(1-r_{xz_1,z_2}^2)(1-r_{yz_2,z_1}^2)}} \tag{7.8}$$

在利用样本研究总体的特性时，由于抽样误差的存在，样本中控制了其他变量的影响，有时可能在样本中两个变量间偏相关系数不为 0，但不能说总体中这两个变量间的偏相关系数不为 0，因此必须进行检验。检验公式为

$$t = r \frac{\sqrt{n-k-2}}{\sqrt{1-r^2}} \tag{7.9}$$

式中， n 为观测量数； k 为控制变量的数目； $n-k-2$ 是自由度。

3. 分析步骤

偏相关分析的步骤可分为两步：

- 第 1 步 根据公式计算偏相关系数。
- 第 2 步 对样本来自的两总体是否存在显著性相关进行推断。

具体如下。

- 提出原假设 H_0 ：即两总体的偏相关系数与零无显著性差异。
- 选择检验统计量：偏相关分析选择的是 t 统计量。
- 计算 t 值及对应的概率 P 值：根据式 (7.9) 计算检验统计量 t 值，同时计算显著性概率 P 值。
- 决策：如果显著性概率 P 值小于给定的显著性水平 α ，则应拒绝原假设，认为两总体的偏相关系数与零有显著性差异；反之，如果检验统计量的显著性概率 P 值大于显著性水平 α ，则不能拒绝原假设，可以认为两总体的偏相关系数与零无显著性差异，即两样本间的偏相关性不显著。

7.3.2 偏相关分析 SPSS 实例分析

【例 7-3】表 7.6 是四川绵阳地区 3 年生中山柏的生长数据，分析月生长量与月平均气温、月降雨量、月平均日照时数、月平均湿度 4 个气候因素中哪些因素有关。（数据来源：袁佳祖，《灰色系统理论》；参见数据文件：data7-3.sav。）

第 1 步 分析。

这 4 个气候因素彼此均有影响，分析时应対生长量与 4 个气候因素分别求偏相关，如在求生长量与气候因素的相关时控制其他因素的影响。然后比较相关系数，按 4 个气候因素对中山柏生长量影响程度的大小排序，需进行偏相关分析。

第 2 步 数据组织。

分别定义变量“month”（月份）、“hgrow”（生长量（cm））、“temp”（月平均气温（℃））、

“rain”（月降雨量（mm））、“hsun”（月平均日照时数）、“humi”（月平均湿度），输入数据并保存。

表 7.6 绵阳地区 3 年生中山柏生长数据

月份	月生长量	月平均气温	月降雨量	月平均日照时数	月平均湿度	月份	月生长量	月平均气温	月降雨量	月平均日照时数	月平均湿度
1	0.01	4.2	17	54.5	81	7	18	24.7	96.9	101.6	83
2	0.5	7.4	10.8	73.8	79	8	19.3	24.5	269.5	164.6	86
3	1.5	10	17.4	84.7	75	9	14.8	22	194.8	81.6	83
4	10.8	16.1	19.7	137	75	10	10.3	18	58.1	84	82
5	13	21.1	248.7	149.6	77	11	8	13.1	4.9	79.3	81
6	16.3	23.9	72.2	109.5	79	12	1	6.8	12.6	66.5	82

第 3 步 进行偏相关分析。

选择菜单“分析→相关→偏相关”，打开如图 7-3 所示的对话框，指定分析变量和控制变量，分析变量“hgrow”和“temp”的偏相关系数，并将“rain”、“hsun”、“humi”设为控制变量。在主对话框中使用系统默认的“双尾”检验，“显示实际显著性水平”，具体设置如图 7-3 所示。



图 7-3 “偏相关性”对话框

第 4 步 主要结果及分析。

运行结果如表 7.7 所示，从中可以看出，月降雨量、月平均日照时数和月平均湿度为控制变量，生长量与月平均气温关系密切，偏相关系数为 0.977，双尾检测的显著性概率为 0.000（表示趋近于 0 的正数），明显小于显著性水平 0.05。故应拒绝原假设，说明中山柏的生长量与气温间存在显著的相关性。

表 7.7 偏相关性检验结果

相关性				
控制变量			月平均气温	生长量
月降雨量 & 月平均日照时数 & 月平均湿度	月平均气温	相关性	1.000	.977
		显著性（双尾）	.	.000
		自由度	0	7
	生长量	相关性	.977	1.000
		显著性（双尾）	.000	.
		自由度	7	0

7.4 距离分析

7.4.1 基本概念及统计原理

1. 基本概念

距离分析是对观测量之间（变量之间）相似或不相似程度的一种测量，是计算一对观测量之间（一对变量之间）的广义距离。这些相似性或距离测量可以用于其他分析过程，例如因子分析、聚类分析或多维定标分析，有助于分析复杂的数据集。例如，是否可以根据汽车的一些特性，如发动机的大小、MPG（每加仑汽油所行驶的距离）和马来来测量两种汽车的相似性？通过计算汽车间的相似性，可以对这些汽车获得一些认识，如哪些汽车彼此类似，哪些彼此不同，还可以考虑对相似性使用分层聚类或多元定标分析去探测深层结构。

2. 统计原理

距离测量又分为非相似性测量和相似性测量两种。

(1) 非相似性测量

① 对定距数据的非相似性（距离）测量可以使用的统计量有：欧氏距离（Euclidean distance）、平方欧氏距离（Squared Euclidean Distance）、切比雪夫距离（Chebychev）、块（Block）距离、明科夫斯基距离（Minkowski）等。

② 对定序数据，主要使用卡方测量（Chi-Square measure）和 Phi 平方测量（Phi-Square measure）。

③ 对二值（只有两种取值）数据变量之间的距离描述，使用欧氏距离、平方欧氏距离、大小差、模式差、形状、方差、兰斯-威廉姆斯等距离统计量。

(2) 相似性测量

两变量之间可以定义相似性测量统计量，用来对两变量之间的相似性进行数量化描述。又分为以下两种：

① 对于定距数据主要使用皮尔逊（Pearson）相关系数和夹角余弦（Cosine）距离。

② 对于二元数据的相似性测量主要包括拉塞尔-拉奥（Russell-Rao）、简单匹配系数（Simple matching）、杰卡德（Jaccard）相似性指数、哈曼（Hamann）相似性测量等 20 余种。

距离又分为个案（观测记录）之间的距离和变量之间的距离两种。

距离分析中不存在假设检验问题，主要是通过 SPSS 自动计算变量或个案之间的相似性或不相似性距离，根据其计算距离值的大小来确定变量或个案之间的相似性或不相似性的强弱。

7.4.2 距离分析 SPSS 实例分析

【例 7-4】已知我国四城市 2004 年各月的日照时数如表 7.8 所示，请分析各城市日照数是否近似。（数据来源：《2005 年中国统计年鉴》，中国统计出版社；参见数据文件：data7-4.sav。）

表 7.8 2004 年四个城市各月份的日照时数

月份	北京	天津	石家庄	大连	月份	北京	天津	石家庄	大连
1	194.7	161.7	193.8	163.5	7	203.2	179.5	185.4	228.5
2	213.5	185.2	219.2	195.3	8	187.4	149.8	152.1	174
3	243.6	166.8	220.9	223.1	9	198.9	178.7	203.4	202.7
4	248.2	214.3	240.9	276.9	10	225.2	194.7	220.7	228.4
5	253.3	221	277.9	243.4	11	201.4	172.8	197.5	172.9
6	202	182.5	213.4	190	12	144	119.1	97.9	167

第1步 分析。

这是4个城市的日照时数是否相似的问题，可用距离分析法实现，既可以计算其相似性测量，也可以计算其不相似性测量。

第2步 数据组织。

分别定义变量“月份”（用字符型变量）、“北京”、“天津”、“石家庄”、“大连”，输入数据并保存。

第3步 设置距离分析主对话框。

选择菜单“分析→相关→距离”，弹出如图7-4所示的“距离”对话框，将4个变量（“北京”、“天津”、“石家庄”、“大连”）移入“变量”框中进行相似性测量计算；在“计算距离”组中选中“变量间”单选框，进行变量间的距离分析；在“测量”单选框组中选中“非相似性”，求解其非相似性测量。以上设置如图7-4所示。

第4步 设置非相似性测量方法。

由于非相似性与相似性测量的方法不同，因此单击“测量（M）...”按钮设置测量方法时会弹出不同的对话框。第2步中设置的测量标准是非相似性，单击“测量（M）...”按钮弹出如图7-5所示的“距离：非相似性测量”对话框。在本例中，“测量”单选框内选择“区间”类型，“测量”统计量选择“欧式距离”计算变量之间的非相似性。



图 7-4 “距离”对话框



图 7-5 “距离：非相似性测量”对话框

图7-5的对话框中提供了3种非相似性测量的测量标准，下面分别给予说明。

（1）区间：对定距数据的非相似性测量选择此类测量标准，单击“区间”单选按钮，其下的“测量”下拉列表框被激活，单击下拉列表框，列出6种可以使用的统计量，这6种统计量的计算公式如表7.9所示。

表 7.9 区间测量统计量计算公式

区间测量标准	公 式
欧氏距离	$d(x, y) = \left[\sum_{i=1}^n (x_i - y_i)^2 \right]^{1/2}$
平方欧氏距离	$d(x, y) = \sum_{i=1}^n (x_i - y_i)^2$
切比雪夫	$d(x, y) = \max_i x_i - y_i $

续表

区间测量标准	公 式
块 (绝对值)	$d(x, y) = \sum_{i=1}^n x_i - y_i $
明可夫斯基	$d(x, y) = \left[\sum_{i=1}^n x_i - y_i ^m \right]^{1/m}, \quad m \text{ 为待定参数}$
定制	$d(x, y) = \left[\sum_{i=1}^n x_i - y_i ^p \right]^{1/q}, \quad p/q \text{ 为待定参数}$

(2) 计数: 计算分类变量的距离测量, 选择该项, “计数”下的“测量”下拉列表框被激活, 单击下拉列表框, 列出两种可以使用的统计量, 其计算公式如表 7.10 所示。

表 7.10 计数测量统计量计算公式

计数测量统计量	公 式
卡方测量	$d_{\text{chi}}(x, y) = \sqrt{\sum_{i=1}^n \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum_{i=1}^n \frac{(y_i - E(y_i))^2}{E(y_i)}}$
Phi 平方测量	$d_{\text{chi}}(x, y) = \frac{d_{\text{chi}}(x, y)}{\sqrt{n}}$

(3) 二元: 计算二元变量的距离测量, 选择该项, “二元”下的“测量”下拉列表框被激活, 单击下拉列表框, 列出 7 种可以使用的统计量。

表 7.11 二元变量列表

1 \ 2	Present	Absent
Present	a	b
Absent	c	d

对二元变量计算距离测量时, 首先建立如表 7.11 所示的列联表。其中 Present 表示该变量具有某些特征。Absent 表示该变量不具有某些特征。a、b、c、d 分别表示满足条件的变量对个数。在 SPSS 中, 默认变量取值为 1 代表 Present, 取值为 0 代表 Absent, 该取值可以通过“二元”下拉列表下方的“存在”框和“不存在”框调整。

“二元”下拉列表主要包括如表 7.12 所示的几项指标。

表 7.12 二分类测量统计量计算公式

统 计 量	公 式
欧氏距离	$d(x, y) = \sqrt{b + c}$
平方欧氏距离	$d(x, y) = b + c$
大小差	$d(x, y) = \frac{(b - c)^2}{(a + b + c + d)^2}$
模式差	$d(x, y) = \frac{bc}{(a + b + c + d)^2}$
方差	$d(x, y) = \frac{b + c}{4(a + b + c + d)}$
形状	$d(x, y) = \frac{(a + b + c + d)(b + c) - (b - c)^2}{(a + b + c + d)^2}$
兰斯-威廉姆斯	$d(x, y) = \frac{b + c}{2a + b + c}$

第 5 步 主要结果及分析。

运行结果如表 7.13 和表 7.14 所示, 具体分析如下。

(1) 数据摘要

表 7.13 所示是变量的个案数及其缺失值情况。

表 7.13 距离非相似性测量个案处理摘要

个案处理摘要					
个案					
有效		缺失		总计	
个案数	百分比	个案数	百分比	个案数	百分比
12	100.0%	0	0.0%	12	100.0%

(2) 距离分析结果表

表 7.14 所示是距离分析的结果表。这是一个对称矩阵，两变量的欧氏距离越大，说明其差别越大，反之越小。从表中可看出“北京”和“大连”的日照数最接近，而“北京”和“天津”的日照数相差最大。表格下方注释说明距离分析采用的是非相似性测量。

表 7.14 距离非相似性测量结果

近似值矩阵				
	欧氏距离			
	北京	天津	石家庄	大连
北京	.000	122.933	71.280	70.542
天津	122.933	.000	111.350	121.427
石家庄	71.280	111.350	.000	110.928
大连	70.542	121.427	110.928	.000
这是非相似性矩阵				

以上例子使用的是非相似性测量方法，如果使用相似性测量，则在第 4 步设置测量方法时，单击“测量(M)…”按钮，会弹出图 7-6 所示的“相似性测量”的设置对话框，注意，此对话框的测量标准与非相似性测量标准有所不同，相似性测量有以下两种测量标准。

- 区间：计算定距变量的相似性测量，主要包括“皮尔逊相关性”和“余弦”两项统计量供选择。
- 二元：计算二元变量的相似性测量。在 SPSS 中，共有 20 种二元变量的相似性测量方法，这里不再一一讲述。



图 7-6 “距离：相似性测量”对话框

在 SPSS 提供的距离分析中，不论是相似性测量还是非相似性测量，都可以对变量或个案数据进行某种标准化处理，并对结果进行转换，图 7-7 所示的对话框下方的“转换值”组可以对个案或变量进行标准化。可以选择的标准化方法如图 7-7 所示。

- 无：不作数据转换，此项为系统默认选项。
- Z 得分：进行标准 Z 分值转换。
- 范围 -1 到 1：将数据标准化到 0~1 之间，方法是将原来的取值除以全距（最大值和最小值之差），如果全距为 0，则所有数据变为 0.5。
- 范围 0 到 1：将数据标准化到 0~1 之间，方法是将原来的取值减去最小值除以全距（最大值和最小值之差），如果全距为 0，则所有数据变为 0.5。

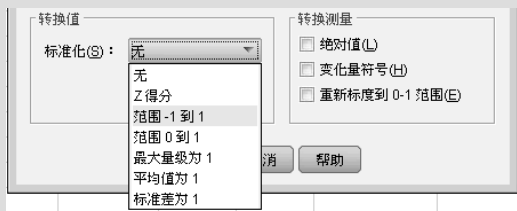


图 7-7 标准化方法

- 最大量为 1：将数据标准化后使其最大值为 1，方法是将原来的取值除以最大值。如果最大值为 0，则值保持不变。
- 平均值为 1：将数据标准化后使其均值为 1，方法是将原来的值除以均值。如果均值为 0，则所有数值加 1。
- 标准差为 1：将数据标准化后使其标准差为 1，方法是将原来的值除以标准差。如果标准差为 0，则值保持不变。

当选择了对数据进行标准化后，还应选择对变量标准化还是对个案标准化。

如果要对距离分析的结果进行转换，则在图 7-6 所示的“转换测量”组中进行设置，共有 3 种转换方法可以选择，每种方法转换之后给出转换的结果。可以同时选中多种方法，得到多种转换结果。

- 绝对值：将结果取绝对值。如果仅仅对距离相关分析的数值大小感兴趣，则可以选择这种方法。
- 变化量符号：改变结果的正负号。
- 重新标度到 0-1 范围：将结果做 0~1 之间的标准化转换。

本例中未对变量做标准化处理，也未对结果进行转换，所以没有做相应的设置，读者如果需要在距离分析时对变量或个案做标准化处理，并对结果进行转换，可参照以上所述进行设置。

☆说明☆

◆ 在选择相似性测量和非相似性测量的方法时，要根据变量类型选择不同的方法。

表 7.15 三个幼崽的数据指标

序号	长	体重	四肢总长	头重
1	50	215	100	11
2	51	220	110	12
3	52	220	112	12

【例 7-5】 某动物产下 3 个幼崽，现分别对 3 个幼崽的长、体重、四肢总长、头重进行测量，试根据这几个测量数据，用距离分析法分析 3 个幼崽的相似性，数据如表 7.15 所示。（参见数据文件：data7-5.sav.）

第 1 步 分析。

这是个案间是否相似的问题，可用距离分析实现，计

算其相似性测量则可以分析出 3 个幼崽是否相似。

第 2 步 数据组织。

建立 5 个变量，分别为“序号”、“长”、“体重”、“四肢总长”和“头重”，录入表 7.15 中的数据即可。

第 3 步 设置距离分析主对话框。

选择菜单：“分析→相关→距离”，弹出“距离”分析的主对话框，进行个案间的相似性分析，其设置如图 7-8 所示。

第 4 步 设置相似性测量方法。

单击图 7-11 “测量”选项组中的“测量(M)…”按钮，弹出如图 7-9 所示的对话框，相似性测量只有两种标准：区间和二元，由于要分析的 4 个变量均为连续型变量，因此选择“区间”中的“皮尔逊相关性”计算个案之间的相似性。由于不对变量进行标准化处理，也不对结果进行转换，所以在“转换值”和“转换测量”两组选项框中未做任何设置。



图 7-8 “距离”对话框设置



图 7-9 相似性测量标准设置

第 5 步 主要结果及分析。

运行结果如表 7.16 和表 7.17 所示，具体分析如下。

(1) 数据摘要

表 7.16 所示是变量的个案数及其缺失值情况，该表说明 3 个个案数据都有效。

(2) 距离分析结果表

表 7.17 列出了 3 个个案之间的相似性分析结果，从表中可以看出，3 个个案（幼崽）的相似性非常高，分别为 0.999 和 1，其中第二个幼崽和第三个幼崽最相似。

表 7.16 距离相似性测量个案处理摘要

个案处理摘要					
个案					
有效		缺失		总计	
个案数	百分比	个案数	百分比	个案数	百分比
3	100.0%	0	0.0%	3	100.0%

表 7.17 距离相似性测量结果

近似值矩阵			
	值的向量之间的相关性		
	1: 1	2: 2	3: 3
1: 1	1.000	.999	.999
2: 2	.999	1.000	1.000
3: 3	.999	1.000	1.000
这是相似性矩阵			

7.5 典型案例

7.5.1 有氧训练中的耗氧量研究

在有氧训练中，人的耗氧量 y （毫升/分 \times 千克体重）是衡量人的身体状况的重要指标，它与年龄 x_1 （岁）、体重 x_2 （千克）、1.5 英里跑所用时间 x_3 （分）、静止时心跳速率 x_4 （次/分）、跑步时心跳速率 x_5 （次/分）、跑步时最大心跳速率 x_6 （次/分）有关。为了研究人的耗氧量与这些变量之间的关系，美国北卡罗莱纳州立大学的健身中心对 31 名测试者进行了测试，得到的数据如表 7.18 所示，以人的耗氧量 y 为因变量， x_1 、 x_2 、 x_3 、 x_4 、 x_5 、 x_6 为自变量，分析因变量和自变量的相关关系。（数据来源：谭荣波等，《SPSS 统计分析实用教程》，科学出版社；参见数据文件：data7-6.sav。）

表 7.18 耗氧量及其相关数据

序号	x_1	x_2	x_3	x_4	x_5	x_6	y
1	44	89.47	11.37	62	178	182	44.609
2	40	75.07	10.07	62	185	185	45.313
3	44	85.84	8.65	45	156	168	54.297
4	42	68.15	8.17	40	166	172	59.571
5	38	89.02	9.22	55	178	180	49.874
6	47	77.45	11.63	58	176	176	44.811
7	40	75.98	11.95	70	176	780	45.681
8	43	81.19	10.85	64	162	170	49.091
9	44	81.42	13.08	63	174	176	39.442
10	38	81.87	8.63	48	170	186	60.055
11	44	73.03	10.13	45	168	168	50.514
12	45	87.66	14.03	56	186	192	37.338
13	45	66.45	11.12	51	176	176	44.754
14	47	79.15	10.6	47	162	164	47.273
15	54	83.12	10.33	50	166	170	51.855
16	49	81.42	8.95	44	180	185	49.156
17	51	69.63	10.95	57	168	172	40.836
18	51	77.91	10	48	162	168	46.672
19	48	91.63	10.25	48	162	164	46.774
20	49	73.37	10.08	67	168	168	50.388

续表

序号	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	y
21	57	73.37	12.63	58	174	176	39.407
22	54	79.38	11.17	62	156	165	46.08
23	52	76.32	9.63	48	164	166	45.441
24	50	70.87	8.92	48	146	155	54.625
25	51	67.25	11.08	48	172	172	45.118
26	54	91.63	12.88	44	168	172	45.118
27	51	73.71	10.47	59	186	188	45.79
28	57	59.08	9.93	59	148	155	50.545
29	49	76.32	9.4	56	186	188	48.673
30	48	61.24	11.5	52	170	176	47.92
31	52	82.78	10.5	53	170	172	47.467

案例分析：要分析因变量和自变量之间的相关关系，可以先做因变量与自变量之间的散点图，然后再根据散点图做二元变量的相关分析，由于数据均为连续型变量，可以通过计算皮尔逊相关系数来衡量二元变量之间的关系。

7.5.2 控制不良贷款

一家大型商业银行在多个地区设有分行，其业务主要是进行基础设施建设、国家重点项目建设、固定资产投资等项目的贷款。近年来，随着经济环境的变化，该银行的贷款额平稳增长，但不良贷款额也有较大比例的提高，这给银行业务的发展带来较大压力。为弄清不良贷款形成的原因，银行行长除了对经济环境进行了广泛的调研外，还希望利用银行业务的有关数据做些定量分析，以便找出控制不良贷款的办法。表 7.19 中的数据就是该银行所属的 25 家分行在 2002 年的主要业务数据。（数据来源：袁卫等，《统计学》，高等教育出版社；参见数据文件：data7-7.sav。）

表 7.19 某商业银行所属的 25 家分行 2002 年的主要业务数据

分行编号	不良贷款/亿元	各项贷款余额/亿元	本年累计应收贷款/亿元	贷款项目个数/个	本年固定资产投资额/亿元
1	0.9	67.3	6.8	5	51.9
2	1.1	111.3	19.8	16	90.9
3	4.8	173.0	7.7	17	73.7
4	3.2	80.8	7.2	10	14.5
5	7.8	199.7	16.5	19	63.2
6	2.7	16.2	2.2	1	2.2
7	1.6	107.4	10.7	17	20.2
8	12.5	185.4	27.1	18	43.8
9	1.0	96.1	1.7	10	55.9
10	2.6	72.8	9.1	14	64.3
11	0.3	64.2	2.1	11	42.7
12	4.0	132.2	11.2	23	76.7
13	0.8	58.6	6.0	14	22.8
14	3.5	174.6	12.7	26	117.1
15	10.2	263.5	15.6	34	146.7
16	3.0	79.3	8.9	15	29.9
17	0.2	14.8	0.6	2	42.1
18	0.4	73.5	5.9	11	25.3

续表

分行编号	不良贷款/亿元	各项贷款余额/亿元	本年累计应收贷款/亿元	贷款项目个数/个	本年固定资产投资额/亿元
19	1.0	24.7	5.0	4	13.4
20	6.8	139.4	7.2	28	64.3
21	11.6	368.2	16.8	32	163.9
22	1.6	95.7	3.8	10	44.5
23	1.2	109.6	10.3	14	67.9
24	7.2	196.2	15.8	16	39.7
25	3.2	102.2	12.0	10	97.1

案例分析：行长想知道，不良贷款是否与贷款余额、应收贷款、贷款项目的多少、固定资产投资等因素有关？如果有关系，它们之间是一种什么样的关系？关系强度如何？这一系列问题都可以通过分析不良贷款和贷款余额、应收贷款、贷款项目的多少、固定资产投资之间的相关关系，并计算相关系数来了解它们之间相关的强度来解决。进一步，还可以将不良贷款与其他几个因素之间的关系用一定的数学关系式表达出来，这将用到第 8 章回归分析的内容。

7.5.3 学生身体状况指标的相似性分析

调查得到 19~22 岁年龄组男性城市学生身体状况指标，如表 7.20 所示，试分析身体状况指标之间的相似性。（参见数据文件：data7-8.sav。）

表 7.20 19~22 岁年龄组男性城市学生身体状况指标

编号	身高/cm	坐高/cm	体重/kg	胸围/cm	肩宽/cm	骨盆宽/cm
1	173.28	93.62	60.10	86.73	38.97	27.51
2	172.09	92.83	60.38	87.39	38.62	27.82
3	171.46	92.73	59.74	85.59	38.83	27.46
4	170.08	92.25	58.04	85.92	38.33	27.29
5	170.61	92.36	59.67	87.46	38.38	27.14
6	171.69	92.85	59.44	87.45	38.19	27.10
7	171.46	92.93	58.70	87.06	38.58	27.40
8	171.60	93.28	59.75	88.03	38.68	27.22

案例分析：各身体状况指标之间的相似性可以用测量变量之间距离的相似性的皮尔逊相关系数来衡量。

7.6 思考与练习

1. 什么是两变量之间的线性相关？两个变量间相关系数的取值范围是什么？负相关系数反映的是两个变量间什么样的关系？
2. 对下列各变量，判断它们之间是否存在相关关系，相关系数为正、负还是零？

(1) 每日卡路里的摄入量与体重。

(2) 海拔与平均气温。

(3) 国内生产总值与新生婴儿的死亡率。

(4) 家庭总收入与文化生活的服务支出。

(5) 结婚年龄与受教育程度。

(6) 每日吸烟数量与肺功能。

3. K.K.Smith 在烟草杂交繁殖的花上收集到如表 7.21 所示的数据，要求对以上 3 组数据两两之间进行相关分析，以 0.05 的显著性水平检验相关系数的显著性。（数据来源：苏金明，《统计软件 SPSS 系列应用实践篇》，电子工业出版社；参见数据文件：data7-9.sav.）

表 7.21 K.K.Smith 所调查的长度资料

花瓣长	49	44	32	42	32	53	36	39	37	45	41	48	45	39	40	34	37	35
花枝长	27	24	12	22	13	29	14	20	16	21	22	25	23	18	20	15	20	13
花萼长	19	16	12	17	10	19	15	14	15	21	14	22	22	15	14	15	15	16

4. 试确定 1962~1988 年安徽省国民收入与城乡居民储蓄存款余额两个变量间的线性相关性，数据如表 7.22 所示。（数据来源：《数据统计与管理》1990 年第 5 期，中国现场统计研究会主办；参见数据文件：data7-10.sav.）

表 7.22 1962~1988 年安徽省国民收入数据表

年 份	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
国民收入（亿元）	34.61	35.67	39.52	47.32	54.14	50.86	49.69	51.61	65.06	72.57	77.72	83.57	82	87.44
存款余额（亿元）	0.59	0.71	0.85	1	1.22	1.14	1.32	1.28	1.35	1.6	1.87	2.2	2.55	2.61
年 份	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	
国民收入（亿元）	95.63	97.23	103.81	116.29	127.87	150.29	161.47	180.2	221.17	271.81	310.53	357.86	444.78	
存款余额（亿元）	2.74	3.13	3.91	5.75	8.76	12.19	16.36	20.95	28.32	38.43	55.43	75.2	89.83	

5. 研究者测量得到 20 名男童身高 X (cm)、体重 Y (kg)、肺活量 Z (L) 的数据如表 7.23 所示，试对控制身高后的体重与肺活量之间的关系进行研究。（数据来源：武松 《SPSS 统计分析大全》，清华大学出版社；参见数据文件：data7-11.sav.）

表 7.23 20 名男童身高、体重、肺活量数据

编号	身高 X	体重 Y	肺活量 Z	编号	身高 X	体重 Y	肺活量 Z
1	139.1	30.4	2	11	139.1	30.4	2
2	163.6	46.2	2.75	12	163.6	46.2	2.75
3	156.2	37.1	2.75	13	156.2	37.1	2.75
4	156.4	35.5	2	14	156.4	35.5	2
5	149.7	31	1.5	15	149.7	31	1.5
6	145	33	2.5	16	145	33	2.5
7	135	27.6	1.25	17	135	27.6	1.25
8	153.3	41	2.75	18	153.3	41	2.75
9	152	32	1.75	19	152	32	1.75
10	160.5	47.2	2.25	20	160.5	47.2	2.25

6. 某高校抽样 10 名短跑运动员，测出 100 米短跑的名次和跳高的名次如表 7.24 所示，问这两个名次是否在 0.05 的显著性水平下具有相关性。（数据来源：马庆国，《应用统计学：数据统计方法、数据获取与 SPSS 应用》，科学出版社；参见数据文件：data7-12.sav.）

表 7.24 10 名运动员的 100 米短跑及跳高名次

百米名次	1	2	3	4	5	6	7	8	9	10
跳高名次	4	3	1	5	2	7	10	8	9	6

7. 某公司太阳镜销售情况如表 7.25 所示，请分析销售量与平均价格、广告费用和日照时间之间的关系，并说明此题用偏相关分析是否有实际意义（显著性水平为 0.05）。（数据来源：卢纹岱，《SPSS for Windows 统计分析（第 3 版）》，电子工业出版社；参见数据文件：data7-13.sav。）

表 7.25 某公司销售太阳镜的数据

月份	1	2	3	4	5	6	7	8	9	10	11	12
销量	75	90	148	183	242	263	278	318	256	200	140	80
价格	6.8	6.5	6	3.5	3	2.9	2.6	2.1	3.1	3.6	4.2	5.2
广告费用	2	5	6	7	22	25	28	30	22	18	10	2
日照时间	2.4	4	5.2	6.8	8	8.4	10.4	11.5	9.6	6.1	3.4	2

第 8 章 回 归 分 析

描述事物数量特征的变量之间存在的主要关系有两种：一种是相关关系，另一种是函数关系。相关关系用相关分析处理，而函数关系一般用回归分析进行研究。相关分析与回归分析都是研究变量之间存在的相互关联关系的方法，但两者之间存在以下三点区别：相关分析研究的变量之间是对等关系，而回归分析研究的变量要区分自变量和因变量；相关分析研究的变量都是随机变量，而回归分析中因变量是随机的，自变量是非随机变量；相关分析只表明现象是否相关、相关方向和密切程度，不能指出变量间相互关系的具体形式，而回归分析可以通过一个数学模型来表现变量之间相关的具体形式。

本章主要介绍回归分析的基本概念及常用的回归分析方法：线性回归分析、曲线回归分析、非线性回归分析和二元 Logistic 回归分析。

8.1 回归分析简介

8.1.1 回归分析的概念

回归分析的基本思想和方法以及“回归 (Regression)”名称的由来，都要归功于英国统计学家 F.Galton 和他的学生——现代统计学的奠基者之一 K.Pearson，他在研究父母身高与其子女身高的遗传关系时，观察了 1078 对夫妇。以每对夫妇的平均身高作为解释变量 x ，取他们的一个成年子女的身高作为被解释变量 y ，将结果在平面直角坐标系上绘成散点图，发现趋近于一条直线。计算出的回归直线方程为

$$\hat{y} = 33.73 + 0.516x \quad (\text{以英寸为单位, } 1 \text{ 英寸} = 2.54 \text{ 厘米})$$

这种趋势表明，父母身高 x 每增加一个单位时，其成年子女的身高 y 也平均增加 0.516 个单位。这个结果表明，虽然高个子父辈有生出高个子子女的趋势，但父母身高增加一个单位，子女身高仅增加半个单位左右。平均来说，一群高个子父辈的子女们在同龄人中平均仅为略高个子；一群矮个子父辈的子女们在同龄人中平均仅为略矮个子，即父辈偏离中心的部分在子代被拉回来一些。正是因为子代的身高有回到同龄人平均身高的这种趋势，才使人类的身高在一定时间内相对稳定，没有出现父辈个子高其子女更高，父辈个子矮其子女更矮的两极分化现象。

这个例子生动地说明了生物学中“种”概念的稳定性。于是 F.Galton 引进了回归 (Regression) 这个词来描述父辈身高 x 与子代身高 y 之间的关系。

回归分析是指通过提供变量之间的数学表达式来定量描述变量间相关关系的数学过程。这一数学表达式通常称为经验公式。我们不仅可以利用概率统计知识，对这个经验公式的有效性进行判定，同时还可以利用这个经验公式，根据自变量的取值来预测因变量的取值。如果是多个因素作为自变量，还可以通过因素分析，找出哪些自变量对因变量的影响是显著的，哪些是不显著的。

回归分析主要解决以下几方面的问题：

➤ 通过大量的数据样本，确定变量之间的函数关系式。

- 对所确定的数学关系式的可信程度进行各种统计检验,并区分出对某一特定变量影响较为显著的变量和不显著的变量。
- 利用所确定的函数关系式,根据一个或几个变量的值来预测或控制另一个特定变量的取值,并给出这种预测或控制的精确度。

8.1.2 回归分析的一般步骤

一个完整的回归分析通常包括以下几步。

第 1 步 确定回归方程中的因变量和自变量。

由于回归分析用于分析一个事物如何随其他事物的变化而变化,因此回归分析的第一步需确定因变量 y 和自变量 x_i 。回归分析正是要建立 x_i 与 y 之间的回归方程,并在给定 x_i 的前提下,通过回归方程预测 y 的取值。

第 2 步 确定回归模型。

根据函数拟合方式,通过观察散点图确定应通过哪种数学模型来概括回归方程。如果被解释变量与解释变量之间存在线性关系,则应进行线性回归分析,建立线性回归模型;反之,如果存在非线性关系,则应进行非线性回归分析,建立非线性回归模型。

第 3 步 建立回归方程。

根据收集到的数据以及第 2 步所确定的回归模型,在一定的统计拟合准则下估计出模型中的各个参数,得到一个确定的回归方程。

第 4 步 对回归方程进行各种检验。

由于回归方程是在样本数据基础上得到的,因此,需要对回归方程进行检验,以确定回归方程是否真实地反映了事物之间的统计关系以及回归方程能否用于预测等。主要包括以下几方面的检验。

(1) 拟合优度检验:检验样本数据聚集在样本回归直线或曲线周围的密集程度,从而判断回归方程对样本数据的代表程度。一般用决定系数 R^2 实现,它越接近于 1,表明回归方程的拟合程度越好;反之,越接近于 0,方程拟合就越差。

(2) 回归方程的显著性检验:对因变量与所有自变量之间的线性关系是否显著的一种假设检验。一般采用 F 检验,其中原假设 H_0 : 回归总体不具显著性(即所有回归系数与零无显著差别: $\beta_0 = \beta_1 = \dots = \beta_p = 0$),备择假设 H_1 : 回归总体具有显著性(即所有自变量对 y 具有显著的线性作用,也就是说,所有回归系数同时与 0 有显著差别)。

(3) 回归系数的显著性检验:根据样本估计的结果对总体回归系数的有关假设进行检验。之所以要对回归系数进行显著性检验,是因为回归方程的显著性检验只能检验所有回归系数是否同时与零有显著性差异,它不能保证回归方程中不包含不能较好地解释因变量变化的自变量,因此,可以通过回归系数显著性检验对每个回归系数进行考察。其中,原假设 H_0 : x_i 对 y 没有显著性影响,备择假设 H_1 : x_i 对 y 具有显著性影响。

第 5 步 利用回归方程进行预测。

建立回归方程的目的之一就是根据回归方程对事物的未来发展趋势进行预测。

以上是进行回归分析的基本步骤,但在处理实际问题时,一定要以问题的专业背景为基础,而不是拘泥于固定的数学方法,这也是统计学与传统数学的根本区别之一。

☆说明☆

- (1) 回归方程的显著性检验旨在检验所有自变量与因变量之间的线性关系是否统计显著,如果线性关系统计显著,说明自变量确实能影响因变量,就可以用自变量的取值去
-

预测因变量的取值；反之则说明自变量与因变量之间没有显著的线性关系。一般采用 F 统计量进行 F 检验， F 检验依赖于 F 分布确定检验临界值，如果计算出的 F 值大于临界值，或者计算出的显著性概率小于 0.05，则说明自变量与因变量之间具有显著的线性关系。

- (2) 回归系数的显著性检验旨在检验单个自变量与因变量之间的线性关系是否统计显著。系数的显著性检验通过 T 检验完成， T 检验依赖于 T 分布计算临界值，如果计算出的 T 值大于临界值或者计算出的显著性概率小于 0.05，说明回归系数具有显著性，单个自变量与因变量之间具有显著的线性关系。
- (3) 在一元线性回归分析中，由于只有一个自变量，所以回归方程的显著性检验可以替代回归系数的显著性检验，但在一般的多元回归条件下，两种检验要说明的问题不同，不能相互替代。

8.2 线性回归分析

8.2.1 基本概念及统计原理

1. 基本概念

线性回归假设因变量与自变量之间为线性关系，用一定的线性回归模型来拟合因变量和自变量的数据，并通过确定模型参数来得到回归方程。根据自变量的多少，线性回归可有不同的划分。当自变量只有一个时，称为一元线性回归；当自变量有多个时，称为多元线性回归。

2. 统计原理

(1) 一元线性回归

如前所述，一元线性回归模型是指只有一个解释变量的线性回归模型，用于表达被解释变量与另一个解释变量之间的线性关系。

一元线性回归的数学模型为

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (8.1)$$

式(8.1)表明， y 的变化可由两部分来解释：第一， x 的变化引起的 y 的线性变化部分，即 $\beta_0 + \beta_1 x$ ；第二，由其他随机因素引起的 y 的变化部分，即 ε 部分。由此可以看出一元线性回归模型是被解释变量与解释变量间非一一对应的统计关系的良好诠释，即当 x 给定后 y 的值并非唯一，但它们之间可以通过 β_0 和 β_1 保持着密切的线性关系。因此，一元线性回归方程如下：

$$E(y) = \beta_0 + \beta_1 x \quad (8.2)$$

式(8.2)表明 x 和 y 之间的统计关系是在平均意义下表述的，即当 x 的值给定后利用回归模型计算得到的 y 值是一个平均值。一元线性回归方程在二维平面上表示为一条直线，表示变量 x 变化时引起变量 y 的变化的估计值。

在实际情况中，某一事物（被解释变量）总会受到多方面因素（多个解释变量）的影响。一元线性回归分析是在不考虑其他影响因素或在认为其他影响因素确定的条件下，分析一个解释变量是如何线性影响解释变量的，因而是比较理想化的分析。

(2) 多元线性回归

多元线性回归模型是指含有多个解释变量的线性回归模型，用于解释被解释变量与其他多个

解释变量之间的线性关系。其数学模型为

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p + \varepsilon \tag{8.3}$$

式 (8.3) 表示一个 p 元线性回归模型, 其中有 p 个解释变量。它表明被解释变量 y 的变化可由两部分组成: 第一, 由 p 个解释变量 x 的变化引起 y 的线性变化部分, 即 $\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$; 第二, 由其他随机因素引起的 y 的变化部分, 即 ε 部分, 叫随机误差。 $\beta_0, \beta_1, \dots, \beta_p$ 都是模型中的未知参数, 分别为回归常数和偏回归系数。则多元线性回归模型的回归方程为

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \tag{8.4}$$

估计多元线性回归方程中的未知参数是多元线性回归分析的核心任务之一。从几何意义上讲, 多元线性回归方程是 p 维空间上的一个超平面, 即回归平面。

8.2.2 一元线性回归 SPSS 实例分析

【例 8-1】 现有 1992~2006 年财政收入和国内生产总值 (单位: 亿元) 的数据, 如表 8.1 所示, 请研究财政收入和国内生产总值之间的线性关系。(数据来源:《中国统计年鉴》, 中国统计出版社; 参见数据文件: data8-1.sav。)

表 8.1 1992 年~2006 年国内生产总值与财政收入数据

年份	国内生产总值	财政收入	年份	国内生产总值	财政收入
1992	26923.5	3483.37	2000	99214.6	13395.23
1993	35333.9	4348.95	2001	109655.2	16386.04
1994	48197.9	5218.10	2002	120332.7	18903.64
1995	60793.7	6242.20	2003	135822.8	21715.25
1996	71176.6	7407.99	2004	159878.3	26396.47
1997	78973.0	8651.14	2005	183867.9	31649.29
1998	84402.3	9875.95	2006	210871.0	38760.20
1999	89677.1	11444.08			

第 1 步 分析。

显然, 财政收入是受国内生产总值影响的, 从经验上看, 二者应该呈线性关系, 这是一个因变量和一个自变量之间的问题, 故考虑用一元线性回归进行分析。

第 2 步 数据组织。

定义 3 个变量, 分别为 “year” (年份)、 “x” (国内生产总值)、 “y” (财政收入), 输入数据并保存。

第 3 步 作散点图, 观察两个变量的相关性。

依次选择菜单 “图形→旧对话框→散点图/点图→简单散点图”, 并将 “国内生产总值” 作为 x 轴, “财政收入” 作为 y 轴, 得到如图 8-1 所示的散点图。可以看出两变量具有较强的线性关系, 可以用一元线性回归来拟合两变量。

第 4 步 一元线性回归分析设置。

(1) 选择菜单: “分析→回归→线性”, 打开 “线性回归” 对话框, 并按如图 8-2 所示进行设置。

(2) “统计” 对话框设置: 单击 “统计 (S) ...” 按钮, 打开 “线性回归: 统计量” 对话框, 并按图 8-3 所示进行设置, 主要由以下几部分组成。

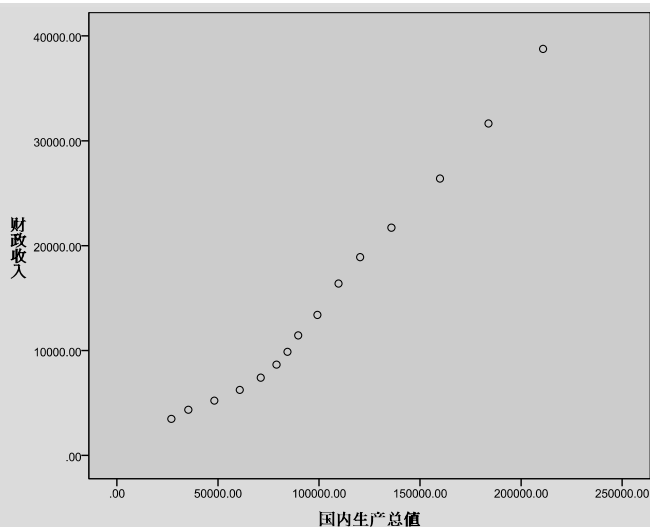


图 8-1 国内生产总值与财政收入的散点图



图 8-2 “线性回归”对话框

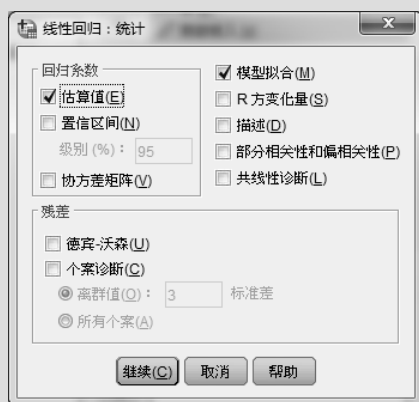


图 8-3 “线性回归: 统计”对话框

- ① “回归系数”复选框组：定义回归系数的输出情况，其中各项的具体作用如下。
- “估算值”选项：输出回归系数的估计值及其标准误差、检验统计量、标准化的回归系数。
 - “置信区间”选项：输出每个回归系数 95% 的置信区间，置信水平是可以设置的。
 - “协方差矩阵”选项：输出每个自变量的相关矩阵、方差、协方差矩阵。
- ② “模型拟合”选项：选中后输出回归模型因变量列表、模型是否恰当的一些检验统计量，以及复相关系数 R 、决定系数 R^2 和调整的 R^2 、方差分析表等。此项为系统默认选项。
- ③ “ R 方变化量”选项：选中后输出模型拟合过程中 R^2 、 F 值和 p 值的改变情况。
- ④ “描述”选项：选中后输出描述性统计量。
- ⑤ “部分相关性和偏相关性”选项：选中后输出自变量的相关系数、部分相关系数和偏相关系数。
- ⑥ “共线性诊断”选项：选中后输出多元线性回归中用于共线性诊断的统计量。
- ⑦ “残差”复选框组：输出残差分析的结果。
- “德宾-沃森”选项：选中后输出 Durbin-Watson 残差序列相关性检验结果。
 - “个案诊断”选项：选中后输出超过规定的 n 倍标准差的残差列表或全部残差列表。
- (3) “图”对话框设置：单击“图(T)...”按钮，打开“线性回归：图”对话框，并按图 8-4 所示进行设置。该对话框主要包括以下几部分。
- ① 候选变量框：列举出可以用来绘制图形的中间统计量，包括因变量 (DEPENDNT)、标准化预测值 (ZPRED)、标准化残差 (ZRESID)、剔除残差 (DRESID)、修正后预测值 (ADJPRED)、用户化残差 (SRESID) 和用户化剔除残差 (SDRESID)。
- ② “散点图 1/1”选项组：从左侧候选变量框中选择变量到 X 、 Y 轴框，定义需要绘制的回归分析诊断图或预测图。
- ③ “标准化残差图”复选框组：选择绘制标准化残差图的类型，包括直方图 (H) 和正态概率图 (R)。
- ④ “生成所有局部图”复选框：选择是否绘制每一个自变量与因变量残差的散点图。

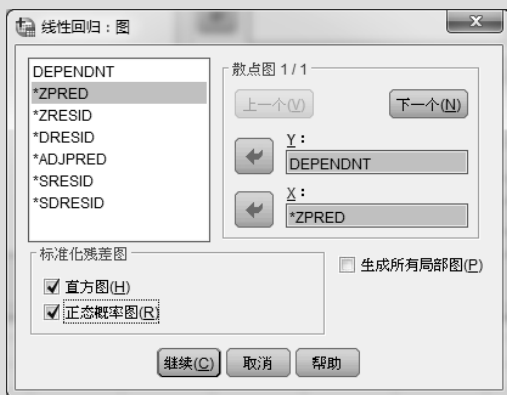


图 8-4 “线性回归：图”对话框

(4) “保存”对话框设置：单击“保存(S)...”按钮，打开“线性回归：保存”对话框，并按图 8-5 所示进行设置。该对话框主要包括以下几部分。

- ① “预测值”复选框组：主要用于保存预测值。
- “未标准化”选项：保存模型对因变量的原始预测值。
 - “标准化”选项：保存标准化后的预测值，此时均值为 0，标准差为 1。

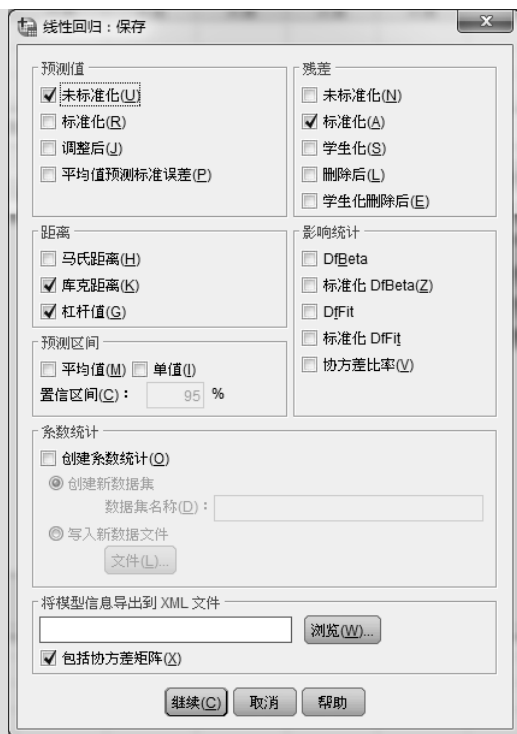


图 8-5 “线性回归：保存”对话框

- “调整后”选项：保存去掉当前记录时，当前模型对该记录因变量的预测值。
- “平均值预测标准误差”选项：保存预测值的标准差。
- ② “残差”复选框组：用于设置残差的保存选项。
 - “未标准化”选项：保存模型预测值对因变量观测值的原始残差。
 - “标准化”选项：保存用 U 变换进行标准化后的残差，此时均值为 0，标准差为 1。
 - “学生化”选项：保存学生化残差，即用 T 变换进行标准化后的残差。
 - “删除后”选项：保存删除当前记录后的残差。
 - “学生化删除后”选项：保存删除当前记录后，用 T 变换进行标准化后的残差。
- ③ “距离”复选框组：保存测量数据点离拟合模型距离的指标，通常用于诊断离群点或强影响点。
 - “马氏距离”选项：保存记录值与样本平均值的马氏（Mahalanobis）距离。
 - “库克距离”选项：保存删除当前记录后，模型残差的变化量。
 - “杠杆值”选项：测量该数据点的影响强度。
- ④ “影响统计”复选框组：保存用于判断强影响点的统计量。
 - “DfBeta”选项：保存去掉该观测值后回归系数的变化量。
 - “标准化 DfBeta”选项：保存标准化的 DfBeta 值，当其大于 n 时，该点可能为强影响点，其中 n 表示样本个数。
 - “DfFit”选项：保存去掉该观测点后预测值的变化值。
 - “标准化 DfFit”选项：保存标准化后的 DfFit 值。
 - “协方差比率”选项：保存去掉该观测点后的协方差阵与含全部观测值的协方差阵的比率。
- ⑤ “预测区间”复选框组：选择是否给出均值和个体参考值的置信区间。

⑥ “系数统计”选项组：主要用于保存上述中间变量。SPSS 23 提供了两种保存方法，可以将结果保存到一个新生成的数据文件中（创建新数据集（A）），也可以将结果直接保存在一个其他的文件中（写入新数据文件（W））。

⑦ “包括协方差矩阵”复选框：选择此项，表示在 XML 文件中保存协方差矩阵。

（5）“选项”对话框设置：单击“选项（O）...”按钮，打开“线性回归：选项”对话框，并按图 8-6 所示进行设置，该对话框主要包括以下几部分。

① “步进法条件”选项：设置变量进入回归模型和排除的标准。

② “在方程中包括常量”选项：用于决定模型中是否包含常数项，默认选中此项。

③ “缺失值”单选项组：定义缺失值的处理方式。

➤ “成列排除个案”选项：只要变量中有数据值缺失就剔除该数据。

➤ “成对排除个案”选项：仅当要分析的变量值缺失时才剔除该数据。

➤ “替换为平均值”选项：用变量均值代替变量缺失值。

第 5 步 主要结果及分析。

运行结果如表 8.2~表 8.6 和图 8-7~图 8-9 所示，分别解释如下。

（1）表 8.2 为变量输入和移去表，表中显示回归模型编号、图 8-6 “线性回归：选项”对话框输入模型的变量、移出模型的变量和变量的筛选方法。从表 8.2 可以看出，输入模型的自变量为“国内生产总值”。表下方的注释含义：a. 因变量是“财政收入”；b.所有的变量均输入回归模型（这里仅一个变量）”。

表 8.2 输入/除去的变量^a表

模型	输入的变量	除去的变量	方法
1	国内生产总值 ^b	.	输入

a. 因变量：财政收入
b. 已输入所请求的所有变量

（2）表 8.3 是模型摘要表，主要是回归方程的拟合优度检验。表中显示相关系数 R、决定系数 R 方、调整后的 R 方和估计值的标准误差等信息，这些信息反映了因变量和自变量之间的线性相关强度。从表 8.3 可看出 $R = 0.989$ ，说明自变量与因变量之间的相关性很强。 $R^2 = 0.979$ ，说明自变量 x 可以解释因变量 y 的 97.9% 的差异性。表下方的注释含义：a.预测变量是“国内生产总值”；b.因变量是“财政收入”。

表 8.3 模型摘要^b表

模型	R	R 方	调整后 R 方	标准估算的误差
1	.989 ^a	.979	.977	1621.66312

a. 预测变量：（常量），国内生产总值
b. 因变量：财政收入

（3）表 8.4 是方差分析表，表中显示因变量的方差来源、方差平方和、自由度、均方、F 检验统计量的观测值和显著性水平。方差来源有回归、残差。从表中可以看出，F 统计量的观测值为 592.25，显著性概率为 0.000，即检验假设“ H_0 : 回归系数 $B = 0$ ”成立的概率为 0.000，从而应拒绝原假设，说明因变量和自变量的线性关系是非常显著的，可建立线性模型。

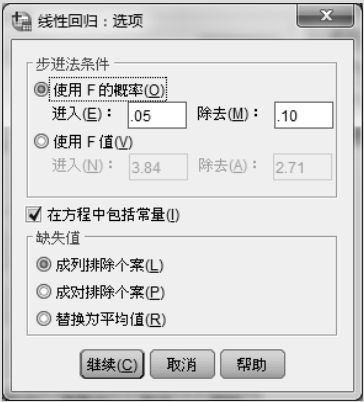


表 8.4 方差分析 (ANOVA^a) 表

模型	平方和	自由度	均方	F	显著性
1 回归	1557492999.819	1	1557492999.819	592.250	.000 ^b
残差	34187286.770	13	2629791.290		
总计	1591680286.589	14			

a. 因变量: 财政收入
b. 预测变量: (常量), 国内生产总值

(4) 表 8.5 是回归系数表, 表中显示回归模型的常数项、非标准化的回归系数 B 值及其标准误差、标准化的回归系数值、统计量 t 值以及显著性水平。从表中可看出, 回归模型的常数项为 -4993.281, 自变量“国内生产总值”的回归系数为 0.197。因此, 可以得出回归方程: 财政收入 = -4993.281 + 0.197 × 国内生产总值。

回归系数的显著性水平为 0.000, 明显小于 0.05, 故应拒绝 T 检验的原假设, 这也说明了回归系数的显著性, 说明建立线性模型是恰当的。

表 8.5 回归系数表

模型	未标准化系数		标准化系数	t	显著性
	B	标准误差	Beta		
1 (常量)	-4993.281	919.356		-5.431	.000
国内生产总值	.197	.008	.989	24.336	.000

a. 因变量: 财政收入

(5) 表 8.6 是残差统计表, 表中依次列出了预测值、标准预测值、预测值的标准误差、调整的预测值、残差、标准残差、学生化残差、剔除残差、学生化剔除残差、马氏距离、库克距离以及居中杠杆值。

表 8.6 残差统计表

	最小值	最大值	平均值	标准偏差	个案数
预测值	315.9509	36589.8320	14925.1933	10547.48785	15
标准预测值	-1.385	2.054	.000	1.000	15
预测值的标准误差	418.964	983.777	570.042	165.910	15
调整后预测值	-494.3054	35325.9648	14789.4147	10469.84809	15
残差	-1928.81250	3167.41895	.00000	1562.67369	15
标准残差	-1.189	1.953	.000	.964	15
学生化残差	-1.239	2.189	.038	1.067	15
剔除残差	-2093.78027	3977.67529	135.77867	1927.94233	15
学生化剔除残差	-1.268	2.646	.084	1.165	15
马氏距离 (D)	.001	4.219	.933	1.181	15
库克距离	.000	.825	.132	.251	15
居中杠杆值	.000	.301	.067	.084	15

a. 因变量: 财政收入

(6) 图 8-7 和图 8-8 是残差分布直方图和观测量累积概率 P-P 图。在回归分析中, 总是假定残差服从正态分布, 这两个图就是根据样本数据的计算结果显示残差分析的实际情况。从残差分布的直方图与附于其上的正态分布曲线的比较, 可以观察出残差分析的正态性。同时, 从观测量累积概率 P-P 图也可以看出残差分布服从正态性。

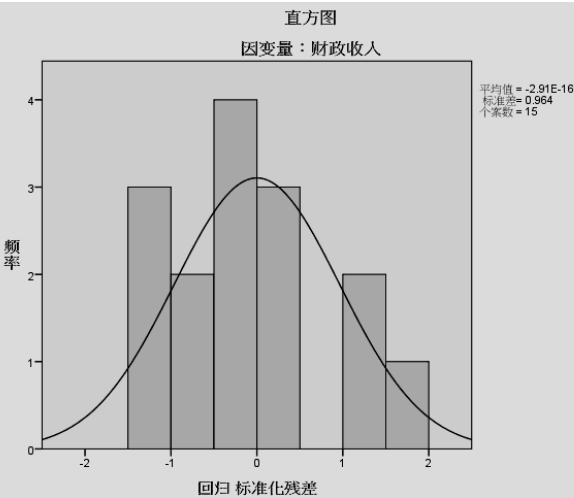


图 8-7 残差分布直方图

(7) 图 8-9 是保存于数据文件中的预测值 (PRE_1)、残差 (RES_1)、库克距离 (COO_1) 和杠杆值 (LEV_1)。

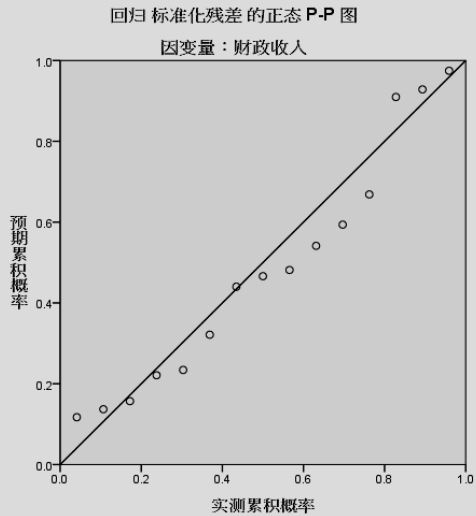


图 8-8 观测量累积概率 P-P 图

year	x	y	PRE_1	RES_1	COO_1	LEV_1
1992	26923.50	3483.37	315.95	3167.4	.61	.14
1993	35333.90	4348.95	1974.46	2374.5	.27	.11
1994	48197.90	5218.10	4511.20	706.9	.02	.07
1995	60793.70	6242.20	6995.05	-752.9	.01	.04
1996	71176.60	7407.99	9042.53	-1634.5	.05	.02
1997	78973.00	8651.14	10579.95	-1928.8	.07	.01
1998	84402.30	9875.95	11650.59	-1774.6	.05	.01
1999	89677.10	11444.08	12690.77	-1246.7	.02	.00
2000	99214.60	13395.23	14571.53	-1176.3	.02	.00
2001	109655.20	16386.04	16630.39	-244.3	.00	.00
2002	120332.70	18903.64	18735.96	167.7	.00	.01
2003	135822.80	21715.25	21790.56	-75.3	.00	.03
2004	159878.30	26396.47	26534.23	-137.8	.00	.09
2005	183867.90	31649.29	31264.90	384.4	.01	.17
2006	210871.00	38760.20	36589.63	2170.4	.83	.30

图 8-9 保存于数据文件中的预测值、残差值等

8.2.3 多元线性回归 SPSS 实例分析

在一元回归中，自变量只有 1 个，因变量只和 1 个因素有关。这在实际情况中是不常见的，常见的情况是 1 个自变量无法将因变量的变化信息完全解释清楚，往往需要多个自变量才能解释因变量的变化信息，这时就会涉及 1 个因变量和 1 组自变量的线性回归问题，这在回归分析中称为多元线性回归。

多元线性回归是为了弥补一元线性回归无法完全解释因变量的变化信息这个缺点而引入的，如果一元线性回归已经很好地说明了因变量的变化，则不必考虑多元线性回归了，只有当一元线性回归效果较差时，才考虑多元线性回归。

【例 8-2】表 8.7 是 1992 年亚洲各国家和地区平均寿命(y)、按购买力计算的人均 GDP(x_1)、成人识字率(x_2)、一岁儿童疫苗接种率(x_3)的数据,试用多元回归的方法分析各国家和地区平均寿命与人均 GDP、成人识字率、一岁儿童疫苗接种率的关系。(数据来源:联合国开发计算署《人类发展报告》,1992 年;参见数据文件: data8-2.sav。)

表 8.7 1992 年亚洲各国和地区的人口现状

序号	国家和地区	y	x_1	x_2	x_3	序号	国家和地区	y	x_1	x_2	x_3
1	日本	79	194	99	99	12	印度尼西亚	62	27	84	92
2	中国香港	77	185	90	79	13	越南	63	13	89	90
3	韩国	70	83	97	83	14	缅甸	57	7	81	74
4	新加坡	74	147	92	90	15	巴基斯坦	58	20	36	81
5	泰国	69	53	94	86	16	老挝	50	18	55	36
6	马来西亚	70	74	80	90	17	印度	60	12	50	90
7	斯里兰卡	71	27	89	88	18	孟加拉国	52	12	37	69
8	中国大陆	70	29	80	94	19	柬埔寨	50	13	38	37
9	菲律宾	65	24	90	92	20	尼泊尔	53	11	27	73
10	朝鲜	71	18	95	96	21	不丹	48	6	41	85
11	蒙古	63	23	95	85	22	阿富汗	43	7	32	35

第 1 步 分析。

这里要分析的是一个变量“平均寿命”与其他三个变量之间的线性关系,显然是一个多元线性回归的问题。

第 2 步 数据组织。

定义 6 个变量,分别为“序号”、“国家和地区”、“ y ”(平均寿命)、“ x_1 ”(人均 GDP)、“ x_2 ”(成人识字率)、“ x_3 ”(一岁儿童疫苗接种率)。输入数据并保存。

第 3 步 多元线性回归分析设置。

(1) 按“分析→回归→线性”顺序打开图 8-2 所示对话框,将变量“ y ”作为因变量,将变量“ x_1 ”、“ x_2 ”、“ x_3 ”作为自变量,将“序号”变量选为“个案标签”。

(2) 在“统计”对话框中选择“估算值”、“模型拟合”、“描述”、“德宾-沃森”。

(3) 在“图”对话框中选择 DEPENDNT 为 Y 轴和 *ZRESID 为 X 轴的散点图。并且选择“直方图”复选框给出正态曲线,选择“正态概率图”复选框输出标准化残差的正态概率图(P-P 图)。

(4) 在“保存”对话框里选择“未标准化”预测值、“未标准化”预测值残差、“标准化”预测值、“标准化”预测值残差。

(5) 在“方法”列表中选择“步进”方法。各种方法的含义解释如下。

- 输入: 默认选项。将自变量框中的自变量全部纳入回归模型中,不做任何筛选。
- 步进: 根据“选项”对话框中设定的条件逐个选取变量进入模型之中。具体选取办法是首先计算各个自变量对因变量的影响大小,选取影响最大的变量进入模型之中,然后重复此过程。注意此时新变量的引入是否会使先前变量丧失统计意义,如果会,这个变量就要剔除并重新计算剩余变量对因变量的影响大小,直到方程中没有可剔除的变量,方程外没有变量可以引入为止。
- 除去: 只出不进,根据移出标准将不进入方程模型的变量一次性全部剔除。
- 后退: 一次性将所有变量引入方程,并依次删除。首先剔除与因变量最小相关且符合剔除标准的变量,然后进行第二个与因变量最小相关且符合剔除标准的变量,依次类推,直到所有变量均符合选入标准为止。

➤ 前进：与向后剔除法相反，首先引入与因变量最大相关且符合引入标准的变量，在引入第一个变量后，再引入第二个与因变量最大偏相关且符合引入标准的变量，依次类推，直到无变量符合引入标准时，终止回归过程。

☆说明☆

◆ 如果要对不同的自变量采用不同的引入法，例如对某两个自变量用“进入”强迫引入法，其他自变量用“逐步”引入法，这时可利用“上一张”与“下一张”按钮把自变量归类到不同的自变量块中，然后对不同的变量子集采用不同的引入方法。

(6) 在“选项”对话框里按默认设置。各选项确认以后提交系统运行。

第 4 步 主要结果及分析。

主要结果如表 8.8~表 8.14 和图 8-10~图 8-12 所示，分别解释如下。

(1)表 8.8 是相关系数矩阵表,显示了包括自变量和因变量在内的 4 个变量的皮尔逊(Pearson)相关系数以及单尾显著性概率。从表中可以看出因变量与自变量的相关系数分别为 0.725、0.847、0.733，单尾检验的显著性概率也较小，说明这三个自变量与因变量的关系均较密切。

表 8.8 相关系数

		平均寿命	人均GDP/100美元	成人识字率（%）	一岁儿童疫苗接种率(%)
皮尔逊相关性	平均寿命	1.000	.725	.847	.733
	人均GDP/100美元	.725	1.000	.503	.307
	成人识字率（%）	.847	.503	1.000	.628
	一岁儿童疫苗接种率（%）	.733	.307	.628	1.000
显著性（单尾）	平均寿命	.	.000	.000	.000
	人均GDP/100美元	.000	.	.009	.082
	成人识字率（%）	.000	.009	.	.001
	一岁儿童疫苗接种率（%）	.000	.082	.001	.
个案数	平均寿命	22	22	22	22
	人均GDP/100美元	22	22	22	22
	成人识字率（%）	22	22	22	22
	一岁儿童疫苗接种率（%）	22	22	22	22

(2)表 8.9 是输入或除去的变量表，系统在进行逐步回归的过程中产生了三个回归模型，模型 1 按照在“选项”对话框确定的标准概率值，先将与平均寿命线性关系最密切的自变量 x_2 (成人识字率) 引入模型，建立 y 与 x_2 之间的一元线性回归模型；模型 2 在此基础上引入了 x_1 (人均 GDP)；模型 3 在模型 2 的基础上引入了变量 x_3 (一岁儿童疫苗接种率)。其中可以看出“逐步”法回归过程中变量被逐步引入的过程。

表 8.9 输入 / 除去的变量^a

模型	输入的变量	除去的变量	方法
1	成人识字率（%）	.	步进(条件: 要输入的 F 的概率 <=.050, 要除去的 F 的概率 >=.100)。
2	人均GDP/100美元	.	步进(条件: 要输入的 F 的概率 <=.050, 要除去的 F 的概率 >=.100)。
3	一岁儿童疫苗接种率（%）	.	步进(条件: 要输入的 F 的概率 <=.050, 要除去的 F 的概率 >=.100)。

a. 因变量：平均寿命

(3)表 8.10 是模型摘要表，分别给出了三个回归模型的复相关系数 R 、决定系数 R 方和调

整后的决定系数 R 方。从第三个模型来看, $R = 0.952$, R 方 = 0.907。从拟合优度上看, 第三个模型明显比第一个模型和第二个模型好。

表 8.10 模型摘要^d

模型	R	R方	调整后R方	标准估算的误差	德宾-沃森
1	.847 ^a	.717	.703	5.501	
2	.915 ^b	.837	.820	4.284	
3	.952 ^c	.907	.891	3.331	1.617

- a. 预测变量: (常量), 成人识字率 (%)
b. 预测变量: (常量), 成人识字率 (%), 人均GDP/100美元
c. 预测变量: (常量), 成人识字率 (%), 人均GDP/100美元, 一岁儿童疫苗接种率 (%)
d. 因变量: 平均寿命

(4) 表 8.11 是方差分析表, 给出了三个模型的方差分析结果。对模型 1: F 值等于 50.628, 显著性概率 P 值为 0.000, 在显著性水平为 0.05 的情形下, 可以认为 y (平均寿命) 与 x_2 (成人识字率) 之间有线性关系。第 2 和第 3 个模型可以进行类似的分析。

表 8.11 方差分析 (ANOVA^d)

模型		平方和	自由度	均方	F	显著性
1	回归	1532.213	1	1532.213	50.628	.000 ^b
	残差	605.287	20	30.264		
	总计	2137.500	21			
2	回归	1788.790	2	894.395	48.733	.000 ^c
	残差	348.710	19	18.353		
	总计	2137.500	21			
3	回归	1937.749	3	645.916	58.205	.000 ^d
	残差	199.751	18	11.097		
	总计	2137.500	21			

- a. 因变量: 平均寿命
b. 预测变量: (常量), 成人识字率 (%)
c. 预测变量: (常量), 成人识字率 (%), 人均GDP/100美元
d. 预测变量: (常量), 成人识字率 (%), 人均GDP/100美元, 一岁儿童疫苗接种率 (%)

(5) 表 8.12 是回归系数表。根据表中数据非标准化系数 B 的数值可知, 逐步回归过程中先后建立的三个回归模型如下。

- 模型 1: $\hat{y} = 38.794 + 0.332x_2$;
模型 2: $\hat{y} = 41.206 + 0.071x_1 + 0.253x_2$;
模型 3: $\hat{y} = 32.993 + 0.072x_1 + 0.169x_2 + 0.179x_3$ 。

同时, 从 t 统计量对应的显著性概率均远小于 0.05 可以判定, 所有回归模型的回归系数都是显著的, 即是有意义的。从表 8.10 的模型拟合优度上看, 显然应以模型 3 作为最终的回归方程。

(6) 表 8.13 是被排除 (剔除) 的变量信息表。表中显示了逐步回归过程中建立的前两个回归模型中剔除的变量信息, 包括各变量的 Beta 值、 t 统计量值、显著性概率、偏相关系数及多重共线性统计量。

(7) 表 8.14 是残差统计表。本表显示了预测值、残差、标准预测值、标准化残差的最小值、最大值、平均值、标准偏差及个案数。根据概率规则的原则, 标准化残差绝对值的最大值为 2.269 < 3, 说明样本数据中没有奇异数据。

表 8.12 回归系数

模型		未标准化系数		标准化系数	t	显著性
		B	标准误差	Beta		
1	(常量)	38.794	3.532		10.983	.000
	成人识字率 (%)	.332	.047	.847	7.115	.000
2	(常量)	41.206	2.825		14.585	.000
	成人识字率 (%)	.253	.042	.645	6.016	.000
	人均GDP/100美元	.071	.019	.401	3.739	.001
3	(常量)	32.993	3.139		10.512	.000
	成人识字率 (%)	.169	.040	.430	4.223	.001
	人均GDP/100美元	.072	.015	.405	4.854	.000
	一岁儿童疫苗接种率 (%)	.179	.049	.339	3.664	.002

a. 因变量：平均寿命

表 8.13 被排除（剔除）的变量

模型		输入 Beta	t	显著性	偏相关	共线性统计
						容差
1	人均GDP/100美元	.401 ^b	3.739	.001	.651	.747
	一岁儿童疫苗接种率 (%)	.333 ^b	2.436	.025	.488	.606
2	一岁儿童疫苗接种率 (%)	.339 ^c	3.664	.002	.654	.606

- a. 因变量：平均寿命
b. 模型中的预测变量：（常量），成人识字率（%）
c. 模型中的预测变量：（常量），成人识字率（%），人均GDP/100美元

表 8.14 残差统计

	最小值	最大值	平均值	标准偏差	个案数
预测值	45.16	81.32	62.50	9.606	22
残差	-7.559	5.301	.000	3.084	22
标准预测值	-1.805	1.959	.000	1.000	22
标准残差	-2.269	1.591	.000	.926	22

a. 因变量：平均寿命

（8）图 8-10 与图 8-11 是残差分布直方图与观测量累积概率 P-P 图。其意义如图 8-7 和图 8-8 的解释。说明残差分布是呈正态分布的。

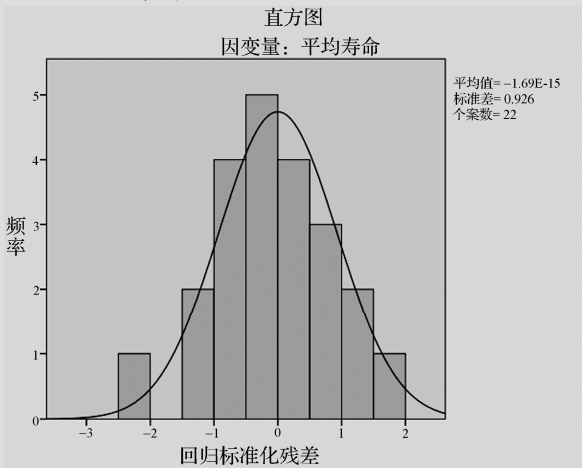


图 8-10 残差分布直方图

(9) 保存于当前数据文件中的预测值 (PRE_1)、残差 (RES_1)、标准化预测值 (ZPR_1)、标准化残差 (ZRE_1)。

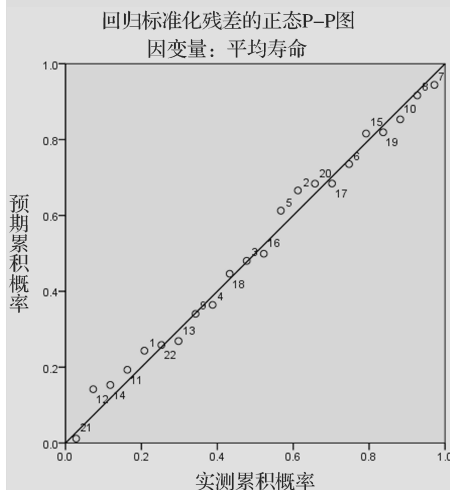


图 8-11 观测量累积概率 P-P 图

...	国家和地区	y	x1	x2	x3	PRE_1	RES_1	ZPR_1	ZRE_1
1	日本	79	194	99	99	81.31619	-2.31619	1.95881	-.69529
2	中国香港	77	185	90	79	75.57225	1.42775	1.36085	.42859
3	韩国	70	83	97	83	70.16440	-.16440	.79788	-.04935
4	新加坡	74	147	92	90	75.15765	-1.15765	1.31769	-.34751
5	泰国	69	53	94	86	68.04678	.95322	.57743	.28614
6	马来西亚	70	74	80	90	67.90477	2.09523	.56265	.62896
7	斯里兰卡	71	27	89	88	65.69915	5.30085	.33304	1.59125
8	中国大陆	70	29	80	94	65.39809	4.60191	.30170	1.38143
9	菲律宾	65	24	90	92	66.36918	-1.36918	.40279	-.41101
10	朝鲜	71	18	95	96	67.49927	3.50073	.52044	1.05087

图 8-12 保存于数据文件中的预测值、残差等 (局部)

☆说明☆

- ◆ 多元线性回归分析的步骤: 选择因变量→确定自变量对因变量的解释力 (一元线性回归)→消除自变量的多重相关性 (步进回归法)→拟合线性回归方程 (多元线性回归)→方程检验 (方程显著性、系数显著性检验)→残差分析→模型确认并用于预测等。在第 2 步, 如果某个自变量能单独说明因变量的大部分信息 (决定系数, R 方在 0.85 以上), 那就直接用一元线性回归拟合方程, 而不必用多元线性回归了。

8.3 曲线回归分析

8.3.1 基本概念及统计原理

1. 基本概念

在实际问题中, 变量间的关系可能是线性的, 也可能是非线性的。若变量间的关系是线性的, 那么可以用线性回归的方法来拟合因变量和自变量之间的关系; 若变量间的关系是非线性的, 问题就复杂得多。变量之间的非线性关系可以划分为本质线性关系和本质非线性关系。所谓本质线性关系是指, 变量关系形式上虽然呈非线性关系 (如二次曲线), 但可通过变量变换化为线性关系, 并可最终通过线性回归分析建立线性模型。本质非线性关系是指, 变量关系不仅形式上呈非线性关系, 而且也无法通过变量变换化为线性关系, 最终无法通过线性回归分析建立线性模型。本节的曲线回归是解决线性关系问题的。

曲线回归 (曲线拟合、曲线估计) 是研究一个自变量和一个因变量之间非线性关系的一种方法。指选定一种用方程表达的曲线, 使得实际数据与理论数据之间的差异尽可能小。如果曲线选择得好, 那么可以揭示因变量与自变量的内在关系, 并对因变量的预测有一定意义。

在曲线回归中, 需要解决两个问题: 一是选用哪种理论模型, 即用哪种方程来拟合观测值; 二是当模型确定后, 如何选择合适的参数, 使得理论数据和实际数据的差异最小。

2. 统计原理

在曲线估计中有很多数学模型, 选用哪种形式的回归方程才能更好地表示出一种曲线的关系, 这往往不是一个简单的问题, 可以用数学方程来表示的各种曲线的数目几乎是没有限量的, 在可能的方程之间, 以吻合度而论, 也许存在着许多吻合同样好的曲线方程。因此, 在对曲线形式的选择上, 对采取什么形式需要有一定的理论, 这些理论是由问题本质决定的。在 SPSS 中, 系统提供了 11 种常见形式的本质线性模型, 如表 8.15 所示。

表 8.15 常见的本质线性模型

模型名称	回归方程	变量变换后的线性方程
线性	$y = b_0 + b_1x$	$y = b_0 + b_1x$
二次方曲线	$y = b_0 + b_1x + b_2x^2$	$y = b_0 + b_1x + b_2x_1 (x_1 = x^2)$
复合曲线	$y = b_0 + b_1^x$	$\ln(y) = \ln(b_0) + \ln(b_1)x$
增长曲线	$y = e^{b_0 + b_1^x}$	$\ln(y) = b_0 + b_1x$
对数曲线	$y = b_0 + b_1 \ln(x)$	$y = b_0 + b_1x_1 (x_1 = \ln(x))$
立方曲线	$y = b_0 + b_1x + b_2x^2 + b_3x^3$	$y = b_0 + b_1x + b_2x_1 + b_3x_2$ ($x_1 = x^2, x_2 = x^3$)
S 曲线	$y = e^{b_0 + b_1/x}$	$\ln(y) = b_0 + b_1x_1 (x_1 = 1/x)$
指数曲线	$y = b_0e^{b_1x}$	$\ln(y) = \ln(b_0) + b_1x$
逆模型	$y = b_0 + b_1 / x$	$y = b_0 + b_1x_1 (x_1 = 1/x)$
幂函数	$y = b_0(x^{b_1})$	$\ln(y) = \ln(b_0) + b_1x_1 (x_1 = \ln(x))$
逻辑函数	$y = \frac{1}{1/\mu + b_0b_1^x}$	$\ln(\frac{1}{y} - \frac{1}{\mu}) = \ln(b_0 + \ln(b_1)x)$

3. 分析步骤

在实际问题中, 用户往往不能确定究竟何种函数模型更接近样本数据, 在 SPSS 中进行曲线估计的一般步骤如下。

- 首先, 根据实际问题的特点, 在上述多种可选择的模型中选择几种;
- 其次, SPSS 自动完成模型参数的估计, 并输出回归方程显著性检验的 F 值和概率 P 值、决定系数 R 方等统计量;
- 最后, 以决定系数为主要依据选择其中的最优模型 (R 方最大的模型), 并进行预测分析。

8.3.2 曲线回归 SPSS 实例分析

【例 8-3】 表 8.16 是 1989~2001 年国家保费收入与国内生产总值的数据, 试研究保费收入与国内生产总值的关系。(资料来源:《中国统计年鉴》, 2002, 中国统计出版社; 参见数据文件: data8-3.sav。)

第 1 步 分析。

先用散点图的形式进行分析, 看究竟是否具有一元线性关系, 如果具有一元线性关系, 则用一元线性回归分析, 否则采用曲线估计求解。

第 2 步 数据组织。

定义为三个变量, 分别是 “year” (年度)、 “y” (保费收入) 和 “x” (国内生产总值), 输入数据并保存。

第 3 步 作散点图，初步判定变量的分布趋势。

依次选择菜单“图形→旧对话框→散点图/点图→简单散点图”，并将“保费收入”作为 y 轴，“国内生产总值”作为 x 轴，得到如图 8-13 所示的散点图。

表 8.16 1989-2001 年保费收入与国内生产总值（单位：亿元）

年度	保费收入	国内生产总值	年度	保费收入	国内生产总值
1980	4.6	4517.8	1991	239.7	21662.5
1981	7.8	4860.3	1992	378	26651.9
1982	10.3	5301.8	1993	525	34560.5
1983	13.2	5957.4	1994	630	46670
1984	20	7206.7	1995	683	57494.9
1985	33.1	8989.1	1996	776	66850.5
1986	45.8	10201.4	1997	1080	73142.7
1987	71.04	11954.5	1998	1247.3	76967.2
1988	109.5	14922.3	1999	1393.22	80579.4
1989	142.6	16917.8	2000	1595.9	88228.1
1990	178.5	18598.4	2001	2109.36	94346.4

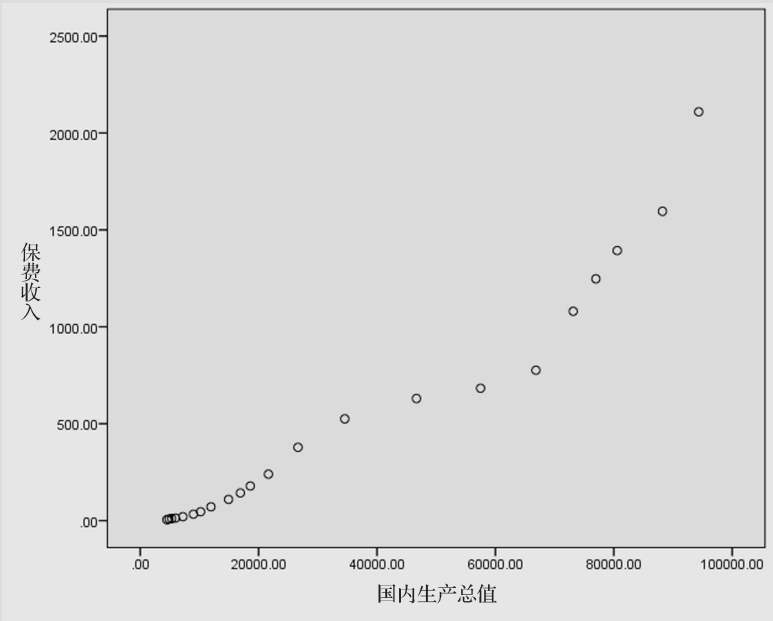


图 8-13 保费收入与国内生产总值的散点图

从图 8-13 可看出，保费收入 y 随国内生产总值 x 的提高而逐渐提高，而且当国内生产总值达到一定水平后，保费收入的增幅更加明显。因此用线性回归模型表示 x , y 的关系是不恰当的。于是应找拟合效果好的模型。

第 4 步 进行曲线回归。

依次选择菜单“分析→回归→曲线估算”，将所有模型全部选上，并按图 8-14 所示设置，看运行结果中哪种模型拟合效果更好（主要看决定系数 R^2 ）。其所有模型的拟合优度 R^2 如表 8.17 所示。



图 8-14 “曲线估算”对话框

从决定系数 (R 方即 R^2) 来看, 三次曲线效果最好 (因为其 R^2 值最大), 并且方差分析的显著性概率值为 0。故重新进行上面的过程, 只选“三次”一种模型。

表 8.17 所有模型的拟合优度值

因变量: 保费收入

方程	模型摘要					参数估算值			
	R 方	F	自由度 1	自由度 2	显著性	常量	b1	b2	b3
线性	.941	316.551	1	20	.000	-154.292	.019		
对数	.772	67.889	1	20	.000	-4576.241	508.979		
逆	.481	18.572	1	20	.000	966.105	-6138735.913		
二次	.973	336.771	2	19	.000	23.846	.003	1.756E-7	
三次	.990	617.659	3	18	.000	-166.430	.029	-5.364E-7	5.022E-12
复合	.789	74.788	1	20	.000	23.315	1.000		
幂	.972	700.929	1	20	.000	2.521E-6	1.796		
S	.946	347.778	1	20	.000	7.069	-27064.140		
增长	.789	74.788	1	20	.000	3.149	5.450E-5		
指数	.789	74.788	1	20	.000	23.315	5.450E-5		

自变量为 国内生产总值。

第 5 步 结果与分析。

主要结果如表 8.18 和图 8-15 所示。

(1) 表 8.18 是对三次曲线模型摘要及参数估算值表, 决定系数 $R^2 = 0.990$, 且显著性概率值为 0.000, 故可判断保费收入与国内生产总值之间有较显著的三次曲线关系。从参数估算值可知因变量与自变量的三次回归模型为: $y = -166.430 + 0.029x - 5.364E - 7x^2 + 5.022E - 12x^3$ 。

表 8.18 三次曲线模型摘要及参数估算值

因变量: 保费收入

方程	模型摘要					参数估算值			
	方	F	自由度 1	自由度 2	显著性	常量	b1	b2	b3
三次	.990	617.659	3	18	.000	-166.430	.029	-5.364E-7	5.022E-12

自变量为 国内生产总值。

(2) 图 8-15 是三次曲线对原始观测值的拟合效果图。可看出其拟合效果非常好。

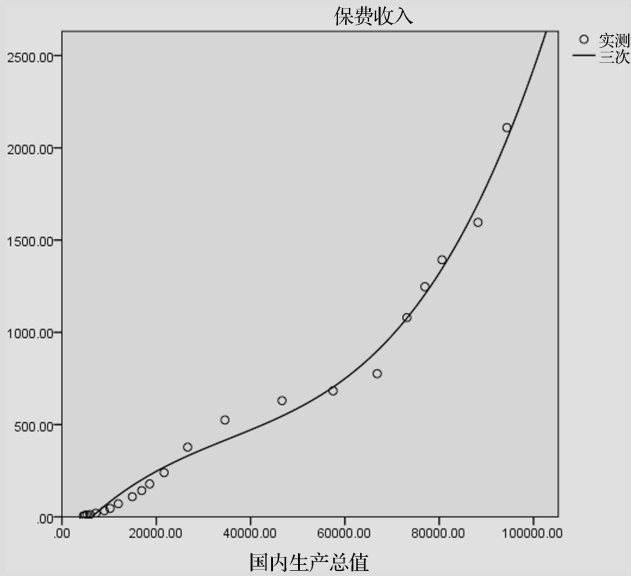


图 8-15 三次曲线的拟合效果图

8.4 非线性回归分析

8.4.1 基本概念及统计原理

1. 基本概念

8.3 节已介绍过，非线性关系可分为本质线性关系和本质非线性关系。可以通过变量转换转化为线性关系，并最终进行线性回归的叫本质线性关系；而无法通过变量转换转化为线性关系，最终也无法进行线性回归分析的叫本质非线性关系。我们平时所讲的非线性回归就是本质非线性关系。

线性回归模型要求变量之间必须是线性关系，曲线回归只能处理能够通过变量转换转化为线性关系的非线性问题，而且也只能用于一个自变量和因变量回归关系的模型分析判别，因此这些方法都有一定的局限性。相反，非线性回归可以估计因变量和自变量之间任意关系的模型，我们可以根据自身需要而设定回归方程的具体形式。因此，非线性回归方法在实际应用中实用价值更大，应用范围更广。

2. 统计原理

非线性回归分析要求自变量和因变量均为数值型变量，如果是分类变量，应该将其重新编码为数值型变量。非线性回归模型一般可以表示为如下形式：

$$y = f(x, \beta) + \varepsilon \tag{8.5}$$

式中， $f(x, \beta)$ 为期望函数，该模型的结构和线性回归模型非常相似，所不同的是期望函数 $f(x, \beta)$ 可能为任意形式，有时甚至可以没有显式表达式。

非线性回归模型的参数估计的基本原理也是先给出一个表示估计误差的函数,即损失函数(为残差绝对值平方和),然后使得该函数取值最小化,并求得此时的参数估计值。其基本思路是:首先为所有未知参数指定一个初始值,然后将原方程按泰勒级数展开,并只取一阶各项作为线性函数的逼近,其余项均归入误差中;然后采用最小二乘法对该模型中的参数进行估计;用参数估计值替代初始值,将方程再次展开,进行线性化,从而又可以求出一批参数估计值。如此反复迭代求解,直到参数估计值收敛为止。

在这一过程中,初始值的设定对模型是否顺利求解影响很大。一个好的初始值不应该偏离真实的参数值太远,否则参数估计的迭代次数可能会增加,或者迭代根本无法收敛,亦或收敛到一个局部最优解而非全局最优解。非线性回归模型在 SPSS 中可以采用 NLR 和 CNLR 两种算法来估计参数,NLR 算法用于寻找能使残差平方和最小的参数估计,CNLR 算法则是首先建立一个非线性的损失函数,然后再寻找最小化这个损失函数的参数估计。常用非线性回归模型的函数形式如表 8.19 所示。

表 8.19 SPSS 提供的非线性回归模型

模型名称	模型表达式
渐近回归	$b_1+b_2*\exp(b_3*x)$
渐近回归	$b_1-(b_2*b_3^x)$
密度	$(b_1+b_2*x)^{(-1/b_3)}$
Gauss	$b_1*(1-b_3*\exp(-b_2*x^2))$
Gompertz	$b_1*\exp(-b_2*\exp(-b_3*x))$
Johnson-Schumacher	$b_1*\exp(-b_2/(x+b_3))$
对数修改	$(b_1+b_3*x)^{b_2}$
对数 Logistic	$b_1-\ln(1+b_2*\exp(-b_3*x))$
Metcherlich 收益递减规律	$b_1+b_2*\exp(-b_3*x)$
Michaelis Menten	$b_1*x/(x+b_2)$
Morgan-Mercer-Florin	$(b_1*b_2+b_3*x^{b_4})/(b_2+x^{b_4})$
Peal-Reed	$b_1/(1+b_2*\exp(-(b_3*x+b_4*x^2+b_5*x^3)))$
三次比	$(b_1+b_2*x+b_3*x^2+b_4*x^3)/(b_5*x^3)$
四次比	$(b_1+b_2*x+b_3*x^2)/(b_4*x^2)$
Richards	$b_1/((1+b_3*\exp(-b_2*x))^{(1-b_4)})$
Verhulst	$b_1/(1+b_3*\exp(-b_2*x))$
Von Bertalanffy	$(b_1^{(1-b_4)}-b_2*\exp(-b_3*x))^{(1/(1-b_4))}$
韦伯	$b_1-b_2*\exp(b_3*x^{b_4})$
产量密度	$(b_1+b_2*x+b_3*x^2)^{(-1)}$

3. 分析步骤

针对呈非线性关系的情况,可以采用两种策略:一是对标准的线性模型做一些修正,使之能处理各种异常情况,但方法仍在线性回归的范畴内,表 8.15 给出了一些常见的非线性回归模型及其变换方式,请读者参照学习;二是彻底打破原有模型的束缚,采用非线性模型来拟合。非线性回归过程是专用的非线性回归模型拟合过程,它采用迭代方法对用户设置的各种复杂曲线模型进行拟合,同时将残差的定义从最小二乘法向外扩展,为用户提供了极为强大的分析能力,不仅能够拟合 SPSS 的回归分析过程提供的全部模型,还可以拟合文件回归、多项式回归、百分位数回归等各种非常复杂的模型。一般第二种策略最权威,同时也是统计学的重点之一,但比较难于掌握。

8.4.2 非线性回归 SPSS 实例分析

【例 8-4】表 8.20 是一家公司在 8 周内每周的营业收入和广告费用数据。公司老板希望建立一个回归模型，以使用电视广告费用和报纸广告费用预测公司营业收入。（资料来源：《SPSS 统计分析大全》，2014，清华大学出版社；参见数据文件：data8-4.sav。）

表 8.20 某公司营业收入与广告费用（单位：万元）

营业收入	电视广告费用	报纸广告费用
96	5	1.5
90	2	2
95	4	1.5
92	2.5	2.5
95	3	3.3
94	3.5	2.3
94	2.5	4.2
94	3	2.5

第 1 步 分析。

这是一个具有两个自变量的模型拟合问题，先用散点图的形式进行分析，看究竟是否具有线性关系，如果具有线性关系，则用线性回归进行分析，否则采用非线性回归求解。

第 2 步 数据组织。

定义 3 个变量，分别是“营业收入”、“电视广告费用”和“报纸广告费用”，输入数据并保存。

第 3 步 作散点图，初步判定变量的分布趋势。

依次选择菜单“图形→旧对话框→散点图/点图→矩阵散点图”，并将 3 个变量均选入到“矩阵变量”框中，运行得到如图 8-16 所示的散点图。

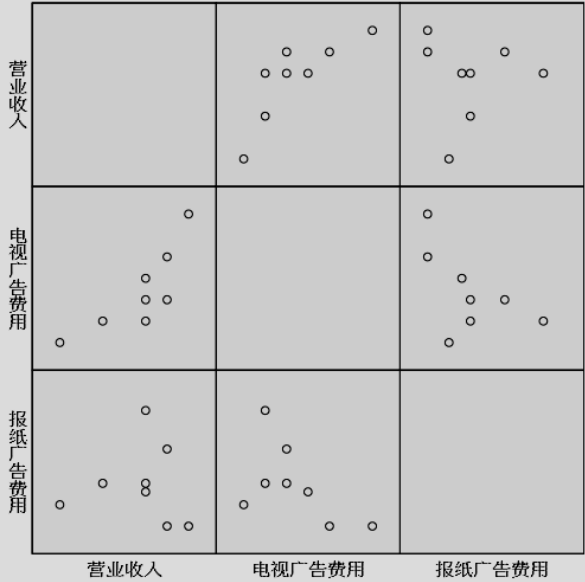


图 8-16 营业收入与广告费用的散点图

图 8-16 所示散点图分为 9 个子图，它们分别描述了三者之间的变化。可以看到，营业收入和两种广告费用存在显著的影响，再观察“电视广告费用”和“报纸广告费用”之间散点图可看到，两种广告费用之间也存在显著的影响关系，说明这两个因变量之间可能存在交叉影响。于是，建立如式 (8.6) 所示的非线性回归模型

$$y = a + bx_1 + cx_2 + dx_1x_2 + \varepsilon \tag{8.6}$$

第 4 步 进行非线性回归。

(1) 选择菜单：“分析→回归→非线性”，打开“非线性回归”对话框，并按如图 8-17 所示进行设置。



图 8-17 “非线性回归”对话框

该对话框主要由以下几部分组成：

- ① 候选变量框：即左上方的变量列表框。
- ② “因变量”框：选择回归模型中的因变量。
- ③ “参数”框：设置模型中所用到的参数并赋初值。
- ④ “模型表达式”框：定义非线性回归模型的表达式。因为非线性回归模型多种多样，所以 SPSS 中直接提供了软键盘和“函数组”来让用户定义非线性模型的表达式。在具体定义的时候，模型中的参数由用户直接通过键盘输入；模型中的变量由候选列表框选入；模型的运算符由用户通过软键盘输入；模型的函数直接从“函数组”框选入。

(2) “参数”设置：单击“参数 (A) ...”按钮，弹出如图 8-18 所示对话框，可以设置参数并逐个添加。如设置不当，也可修改和删除某个参数。本例中，将初始参数值设置为： $a=20$ ， $b=0.2$ ， $c=0$ ， $d=1$ 。

(3) “模型表达式”设置：根据前面分析，通过软键盘将其设为： $a+b*$ 电视广告费用 $+c*$ 报纸广告费用 $+d*$ 电视广告费用*报纸广告费用。

(4) “保存”对话框设置：打开保存对话框，有“预测值”、“残差”、“导数”和“损失函数值”几项，本例选择“预测值”和“残差”两项。

(5) “选项”对话框设置：打开“选项”对话框，按如图 8-19 所示设置。在该对话框中，根据是否选择“标准误差的自助抽样估算”复选框而选择不同的估算方法，并设置迭代情况。

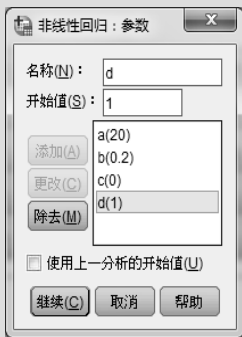


图 8-18 “非线性回归：参数”对话框



图 8-19 “非线性回归：选项”对话框

第 5 步 主要结果及分析。

运行结果如表 8.21、表 8.22 和图 8-20 所示，分别解释如下。

(1) 表 8.21 为迭代历史记录表。可以看出，经过 9 次迭代后，模型达到收敛标准，找到了最佳解。于是得到营业收入关于两种广告费用的预测回归模型为

$$y = 86.531 + 1.089x_1 - 0.667x_2 + 0.724x_1x_2 \quad (8.7)$$

表 8.21 迭代历史记录^b

迭代编号 ^a	残差平方和	参数			
		a	b	c	d
0.1	34498.035	20.000	.200	.000	1.000
1.1	1904.999	20.855	2.938	2.100	7.358
2.1	1120.777	23.716	10.794	1.177	4.096
3.1	757.802	53.915	20.471	40.733	-16.722
4.1	77.094	43.075	15.292	21.727	-6.967
5.1	34.672	57.661	10.188	13.261	-3.935
6.1	23.269	66.613	8.285	11.532	-3.720
7.1	2.609	91.029	-.536	-3.422	1.727
8.1	1.499	86.531	1.089	-.667	.724
9.1	1.499	86.531	1.089	-.667	.724

将通过数字计算来确定导数。

a. 主迭代号在小数点左侧显示，次迭代号在小数点右侧显示。

b. 运行在 9 次迭代后停止。已找到最优的解。

(2) 表 8.22 为整个模型的显著性检验结果。可以看出，决定系数 R^2 为 0.941，拟合结果较好。

表 8.22 方差分析 (ANOVA^a) 表

源	平方和	自由度	均方
回归	70336.501	4	17584.125
残差	1.499	4	.375
修正前总计	70338.000	8	
修正后总计	25.500	7	

因变量：营业收入

a. $R^2 = 1 - (\text{残差平方和}) / (\text{修正平方和}) = .941$ 。

(3) 预测值和残差保存到数据表中，如图 8-20 所示。

营业收入	电视广告费用	报纸广告费用	PRED_	RESID_1
96	5.0	1.5	96.40	-.40
90	2.0	2.0	90.27	-.27
95	4.0	1.5	94.23	.77
92	2.5	2.5	92.11	-.11
95	3.0	3.3	94.76	.24
94	3.5	2.3	94.63	-.63
94	2.5	4.2	94.05	-.05
94	3.0	2.5	93.56	.44

图 8-20 数据表中保存结果

☆说明☆

- (1) 非线性回归中，对模型的初步判断与采用至关重要，需要分析人员具有较好的选择和判断能力。
- (2) 参数初始值的设置会影响迭代过程的收敛性，如果可能的话，应尽量为参数选择合理的、接近于期望的初始值。
- (3) 有时对一个特定的问题，一种算法可能比另一种算法更好。因此可在“选项”对话框里，换用另一种算法。
- (4) 迭代过程终止是因为达到了最大迭代步数，这时得到的这个“最终”模型未必是最好的解，因此可以给参数设置不同的初始值继续进行迭代。
- (5) 如模型需要进行求幂运算或者数据量很大，就可能使计算结果的数值过大或过小而溢出，可以通过适当选取初始值或者利用参数的约束来避免这种情况的发生。

8.5 二元 Logistic 回归分析

8.5.1 基本概念及统计原理

1. 基本概念

前面介绍的线性回归、曲线回归和非线性回归都要求因变量是定量变量，但实际问题中，因变量既有定量的，也有定性的，Logistic 回归分析就是针对因变量是定性变量的回归分析。

在实际生活中，我们经常会遇到因变量是定性变量的情况，如医学上的阴性和阳性，生存与死亡，消费现象中的购买行为发生还是不发生，金融现象中的上市公司 IPO 通过还是不通过，等等。

可以处理定性因变量的统计分析方法有：判别分析、Probit 分析、Logistic 回归分析和对数线性模型分析等。在社会科学中，应用最多的是 Logistic 回归分析。根据因变量取值类别数量不同，Logistic 回归分析又分为二元 Logistic 回归分析和多元 Logistic 回归分析。二元 Logistic 回归模型中因变量只可以取两个值 1 和 0（虚拟因变量），而多元 Logistic 回归模型中因变量可取多个值。本节重点介绍二元 Logistic 回归模型，对于多元 Logistic 回归模型只做简要说明。

2. 统计原理

(1) logit 变换

设因变量 y 是只取 0 或 1 的二分类变量， p 为某事件发生的概率，取值区间为 $[0, 1]$ ，当事件

发生时 $y = 1$ ，否则 $y = 0$ ，即 $p = P(y = 1)$ （事件发生的概率）是研究对象。将比率 $p/(1-p)$ 取自然对数，即对 p 做 logit 变换

$$\text{logit}(p) = \ln(p / (1 - p)) \tag{8.8}$$

当 $p = 1$ 时 $\text{logit}(p) = +\infty$ ，当 $p = 0.5$ 时 $\text{logit}(p) = 0$ ，当 $p = 0$ 时 $\text{logit}(p) = -\infty$ ，故 $\text{logit}(p)$ 的取值范围为 $(-\infty, +\infty)$ 。注意，式中等号右边的分数部分 $p/(1-p)$ 是“事件发生”与“事件不发生”的概率比，称为优势（odd）。所以，logit 变换有很好的统计解释，它是优势的对数。

(2) Logistic 回归模型

设有 k 个因素 x_1, x_2, \dots, x_k 影响 y 的取值，则称

$$\ln(p / (1 - p)) = g(x_1, x_2, \dots, x_k) \tag{8.9}$$

为二维 Logistic 回归模型，简称 Logistic 回归模型，其中的 k 个因素 x_1, x_2, \dots, x_k 称为 Logistic 回归模型的协变量。最重要的 Logistic 回归模型是 Logistic 线性回归模型

$$\ln(p / (1 - p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \tag{8.10}$$

式中， $\beta_0, \beta_1, \dots, \beta_k$ 是待估计的未知参数。可得

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} \tag{8.11}$$

(3) 统计检验

与线性回归一样，拟合时也要考虑模型是否合适、哪些变量该保留、拟合效果如何等问题。线性回归中常用的是决定系数 R^2 ， T 检验、 F 检验等工具在这里均不再适用。在 Logistic 回归中常用的检验有-2 对数似然检验（-2 log（likelihood），2LL）、Hosmer 和 Lemeshow 的拟合优度检验、Wald 检验等。为了简化起见，这里省略相关模型的介绍（读者可参考其他资料），具体检验过程参看下节案例。

8.5.2 二元 Logistic 回归 SPSS 实例分析

【例 8-5】诊断发现运营不良的金融企业是审计核查的一项重要功能，审计核查的分类失败会导致灾难性的后果。表 8.23 列出了 66 家公司的部分运营财务比率，其中 33 家在 2 年后破产（ $y = 0$ ），另外 33 家在同期保持偿付能力（ $y = 1$ ）。请用变量 x_1 （未分配利润/总资产）、 x_2 （税前利润/总资产）和 x_3 （销售额/总资产）拟合一个 Logistic 回归模型。（数据来源：《例解回归分析》，中国统计出版社；参见数据文件：data8-5.sav。）

表 8.23 有偿付能力及破产公司的财务比率

x_1	x_2	x_3	y	x_1	x_2	x_3	y
-62.8	-89.5	1.7	0	43	16.4	1.3	1
3.3	-3.5	1.1	0	47	16	1.9	1
-120.8	-103.2	2.5	0	-3.3	4	2.7	1
-18.1	-28.8	1.1	0	35	20.8	1.9	1
-3.8	-50.6	0.9	0	46.7	12.6	0.9	1
-61.2	-56.2	1.7	0	20.8	12.5	2.4	1
-20.3	-17.4	1	0	33	23.6	1.5	1
-194.5	-25.8	0.5	0	26.1	10.4	2.1	1
20.8	-4.3	1	0	68.6	13.8	1.6	1
-106.1	-22.9	1.5	0	37.3	33.4	3.5	1

续表							
x_1	x_2	x_3	y	x_1	x_2	x_3	y
-39.4	-35.7	1.2	0	59	23.1	5.5	1
-164.1	-17.7	1.3	0	49.6	23.8	1.9	1
-308.9	-65.8	0.8	0	12.5	7	1.8	1
7.2	-22.6	2	0	37.3	34.1	1.5	1
-118.3	-34.2	1.5	0	35.3	4.2	0.9	1
-185.9	-280	6.7	0	49.5	25.1	2.6	1
-34.6	-19.4	3.4	0	18.1	13.5	4	1
-27.9	6.3	1.3	0	31.4	15.7	1.9	1
-48.2	6.8	1.6	0	21.5	-14.4	1	1
-49.2	-17.2	0.3	0	8.5	5.8	1.5	1
-19.2	-36.7	0.8	0	40.6	5.8	1.8	1
-18.1	-6.5	0.9	0	34.6	26.4	1.8	1
-98	-20.8	1.7	0	19.9	26.7	2.3	1
-129	-14.2	1.3	0	17.4	12.6	1.3	1
-4	-15.8	2.1	0	54.7	14.6	1.7	1
-8.7	-36.3	2.8	0	53.5	20.6	1.1	1
-59.2	-12.8	2.1	0	35.9	26.4	2	1
-13.1	-17.6	0.9	0	39.4	30.5	1.9	1
-38	1.6	1.2	0	53.1	7.1	1.9	1
-57.9	0.7	0.8	0	39.8	13.8	1.2	1
-8.8	-9.1	0.9	0	59.5	7	2	1
-64.7	-4	0.1	0	16.3	20.4	1	1
-11.4	4.8	0.9	0	21.7	-7.8	1.6	1

第 1 步 分析。

共有 3 个自变量，均是定量数据类型，而因变量是定性的，取值有两种状态（0 和 1），这是一个典型的可用二元 Logistic 回归解决的问题。

第 2 步 数据组织。

定义三个自变量 x_1 ， x_2 和 x_3 ，再定义因变量 y ，输入数据并保存。

第 3 步 二元 Logistic 回归分析设置。

（1）选择菜单“分析→回归→二元 Logistic”，打开二元“Logistic 回归”对话框，并按图 8-21 所示进行设置。



图 8-21 “Logistic 回归”对话框

(2) “保存”对话框设置: 单击“保存(S)…”按钮, 打开“Logistic 回归: 保存”对话框, 其选项与图 8-5 所示类似, 这里不再赘述。在其中选择“预测值”选项组中的“概率”和“组成员”两项, 即将预测的概率和分类保存下来。

(3) “选项”对话框的设置: 单击“选项(O)…”按钮, 打开“Logistic 回归: 选项”对话框, 按图 8-22 所示进行设置, 并单击“继续”按钮, 返回上一个对话框, 然后单击“确定”按钮, 即可得到分析结果。



图 8-22 “Logistic 回归: 选项”对话框

- ① “统计和图”选项组: 其中的选项用来选择输出哪些统计量或统计图表, 具体选项如下。
- “分类图”选项: 通过比较因变量的观测值和预测值之间的关系, 反映回归模型的拟合效果。
 - “霍斯默-莱梅肖拟合优度”选项: 检验整个回归模型的 Hosmer-Lemeshow 拟合优度。
 - “个案残差列表”选项: 输出标准方差大于某值的个案或全部个案的入选状态, 以及因变量的观测值和预测值及其相应的预测概率、残差值。
 - “估算值的相关性”选项: 输出模型中各估计参数间的相关矩阵。
 - “迭代历史记录”选项: 输出参数估计迭代过程中的系数及对数似然值。
 - “Exp(B) 的置信区间”选项: 选中该选项将会在模型检验的输出结果中列出 Exp(B) (各回归系数指数函数值) 的 $N\%$ (默认值为 95%) 的置信区间, 如果要改变默认值, 可以在空白方框内输入 1~99 之间的任何一个整数。
- ② “显示”单选项组用来选择输出计算结果的方式, 具体说明如下。
- “在每个步骤”选项: 显示 SPSS 每个步骤的计算结果。
 - “在最后一个步骤”选项: 只显示最终计算结果。
- ③ “步进概率”框用来设定步长标准, 以便逐步控制自变量进入方程或被剔除方程。
- “进入”框: 设置变量进入方程的标准值。如果变量的分数统计概率小于所设置进入方程的标准值, 则该变量进入模型, SPSS 默认的显著性水平为 0.05。
 - “除去”框: 设置变量被剔除方程的标准值。如果变量的分数统计概率大于所设置被剔除方程的标准值, 则该变量被剔除方程, SPSS 默认的显著性水平为 0.10。

☆说明☆

- ◆ 进入值应小于删除值, 否则自变量一进入方程就会立即被剔除了。

④“分类分界值”框：用以设置个案分类的断点值。因变量预测值大于分类中止点的个案设为一类，小于分类中止点的个案设为另一类，默认值为0.5，当然也可以重新设置。

⑤“最大迭代次数”框：用以确定达到最大对数似然值之前的迭代次数。最大对数似然值是通过反复迭代计算直到收敛为止而得到的，默认的最大迭代次数为20，当然也可以重新设置。

⑥“在模型中包括常量”选项：用以确定所求模型的参数是否要包含常数项。

(4)“方法”下拉列表框：用以选择自变量进入模型的方法，主要有以下三种。

- “输入”法：所有自变量都强行进入回归模型。
- “向前（有条件/LR/瓦尔德）”：依据条件参数似然比检验结果/偏似然比检验结果/Wald检验结果剔除变量的向前剔除法。
- “向后逐步（有条件/LR/瓦尔德）”：依据条件参数似然比检验结果/偏似然比检验结果/Wald检验结果剔除变量的向后剔除法。

第4步 主要结果及分析。

运行结果如表8.24~表8.32和图8-23所示，分别解释如下。

(1)表8.24是个案处理摘要信息，给出了数据进入模型的记录数。

表 8.24 个案处理摘要

未加权个案 ^a		个案数	百分比
选定的个案	包括在分析中的个案数	66	100.0
	缺失个案数	0	.0
	总计	66	100.0
未选定的个案		0	.0
总计		66	100.0

a. 如果权重处于生效状态，请参阅分类表以了解个案总数。

(2)表8.25是因变量的赋值表，在SPSS中，默认将二分类变量中出现次数较多的赋值为1。本例比较特殊，二分类变量的两种情况出现的次数是一样的，从表格中可以看出，将“两年后破产”赋值为0，“两年后仍有偿付能力”赋值为1。

表 8.25 因变量编码

原值	内部值
两年后破产	0
两年后仍有偿付能力	1

(3)表8.26是模型初始分类预测表，此时模型中不含任何自变量，只包含常数项。表格左方代表实际观测值，右方代表模型的预测值和正确率。此时预测所有公司在两年后仍有偿付能力，预测的正确率为50%。

表 8.26 分类表^{a,b}

	实测		预测		
			y		正确百分比
			两年后破产	两年后仍有偿付能力	
步骤0	y	两年后破产	0	33	.0
		两年后仍有偿付能力	0	33	100.0
	总体百分比				50.0

a. 常量包括在模型中。

b. 分界值为 .500

(4) 表 8.27 和表 8.28 给出了模型系数的检验结果，其中常数项系数为 0.000，其显著性概率为 1，可见常数项不显著。 x_1 、 x_2 和 x_3 的相伴概率分别是 0.000，0.000 和 0.094，如果以 5% 为置信的话， x_1 和 x_2 的系数是显著的。

表 8.27 方程中的变量

		B	标准误差	瓦尔德	自由度	显著性	Exp (B)
步骤 0	常量	.000	.246	.000	1	1.000	1.000

表 8.28 未包括在方程中的变量

			得分	自由度	显著性
步骤 0	变量	x1	31.621	1	.000
		x2	19.358	1	.000
		x3	2.800	1	.094
	总体统计		37.613	3	.000

(5) 表 8.29 是模型系数的 Omnibus 检验结果。共采用了三种检验方法，分别是步与步间的相对似然比检验、块 (Block) 间的相对似然比检验和模型间的相对似然比检验。由于本例中只有一个自变量组且采取强行进入法将所有变量纳入模型，所以三种检验方法的结果是一致的，模型有显著的统计意义。

表 8.29 模型系数的 Omnibus 检验结果

		卡方	自由度	显著性
步骤1	步骤	85.683	3	.000
	块	85.683	3	.000
	模型	85.683	3	.000

(6) 表 8.30 是模型情况摘要表。主要给出了对数似然值的两个决定系数，从数据上看，模型的拟合度不错。

表 8.30 模型摘要

步骤	-2 对数似然	考克斯-斯奈尔R方	内戈尔科R方
1	5.813a	.727	.969

a. 由于参数估算值的变化不足 .001，因此估算在第 12 次迭代时终止。

(7) 表 8.31 是模型的分类预测情况表。此时模型的预测准确率已达到 97%。

表 8.31 分类表^a

		预测		
		y		正确百分比
		两年后破产	两年后仍有偿付能力	
步骤1	实测			
	y	两年后破产	32	1
		两年后仍有偿付能力	1	32
总体百分比				97.0

a. 分界值为 .500

(8) 表 8.32 是 Logistic 模型的拟合结果。表格从左到右依次表示变量及常数项的系数值 (B)、标准误差 (S.E.)、瓦尔德 (Wald) 卡方值、自由度 (df)、显著性概率、Exp (B)。由于各回归

系数均为正数, 取相应的指数后会大于 1, 表示 x_1, x_2 和 x_3 的取值越大, “两年后具有偿付能力”的可能性比“两年后破产”的可能性就越大, 其 Logistic 回归模型为

$$\ln(p/(1-p)) = 0.331x_1 + 0.181x_2 + 5.087x_3 - 10.153$$

则有

$$p = \frac{e^{(0.331x_1 + 0.181x_2 + 5.087x_3 - 10.153)}}{1 + e^{(0.331x_1 + 0.181x_2 + 5.087x_3 - 10.153)}}$$

若预测值 p 的概率小于 0.5, 则样本被归于“两年后破产”组, 相对的, 若预测值 p 的概率大于 0.5, 则样本被归于“两年后有偿付能力”组, 其预测结果值 (局部) 如图 8-23 所示, 其中 PRE_1 表示预测概率值, PGR_1 表示预测分类结果值。

表 8.32 方程中的变量

		B	标准误差	瓦尔德	自由度	显著性	Exp (B)	EXP (B) 的95%置信区间	
								下限	上限
步骤 1 ^a	x1	.331	.301	1.213	1	.271	1.393	.772	2.511
	x2	.181	.107	2.862	1	.091	1.198	.972	1.478
	x3	5.087	5.082	1.002	1	.317	161.979	.008	3430718.695
	常量	-10.153	10.840	.877	1	.349	.000		

a. 在步骤 1 输入的变量: x1, x2, x3。

x1	x2	x3	y	PRE_1	PGR_1
-62.80	-89.50	1.70	0	.00000	0
3.30	-3.50	1.10	0	.01635	0
-120.80	-103.20	2.50	0	.00000	0
-18.10	-28.80	1.10	0	.00000	0
-3.80	-50.60	.90	0	.00000	0
-61.20	-56.20	1.70	0	.00000	0
-20.30	-17.40	1.00	0	.00000	0
-194.50	-25.80	.50	0	.00000	0
20.80	-4.30	1.00	0	.74004	1
-106.10	-22.90	1.50	0	.00000	0
-39.40	-35.70	1.20	0	.00000	0
-164.10	-17.70	1.30	0	.00000	0
-308.90	-65.80	.80	0	.00000	0
7.20	-22.60	2.00	0	.15692	0
-118.30	-34.20	1.50	0	.00000	0

图 8-23 数据文件中保存的结果图

☆说明☆

◆ 从表 8.32 中瓦尔德 (Wald) 检验的显著性概率值可知, 各变量及常数的系数都没有显著的统计意义, 说明这个拟合模型并不是最好的, 读者可使用“向前”或“向后”的逐步方法再行尝试。

8.6 典型 案 例

8.6.1 水稻产量影响因素分析

为了研究水稻产量所受因素的影响, 某研究机构记录了某地区近 18 年的水稻产量、化肥使用量、生猪存栏数、水稻扬花期降雨量的数据, 数据中含有 18 个观测样本, 代表 18 个年份, 有 7 个属性变量: Id (序号)、 x_1 (水稻播种面积)、 x_2 (化肥使用量)、 x_3 (生猪存栏数)、 x_4 (水稻

扬花期降雨量)、 y (水稻总产量) 及 $year$ (年份), 具体数据如表 8.33 所示。(数据来源: 吕振通等,《SPSS 统计分析与应用》,机械工业出版社;参见数据文件: data8-6.sav。)

表 8.33 某地区近 18 年水稻产量数据

Id	x_1	x_2	x_3	x_4	y	year	Id	x_1	x_2	x_3	x_4	y	year
1	147	2	15	27	154.5	1993	10	155	18	51	22	270.5	2002
2	148	3	26	38	200	1994	11	156	23	53	39	298.5	2003
3	154	5	33	20	227.5	1995	12	155	23.5	51	28	229	2004
4	157	9	38	33	200	1996	13	157	24	51	46	309.5	2005
5	153	6.5	41	43	208	1997	14	156	30	52	59	309	2006
6	151	5	39	33	229.5	1998	15	159	48	52	70	371	2007
7	151	7.5	37	46	265.5	1999	16	164	95.5	57	52	402.5	2008
8	154	8	38	78	229	2000	17	164	93	68	36	429.5	2009
9	155	13.5	44	52	303.5	2001	18	156	97.5	74	37	427.5	2010

案例分析: 首先研究水稻产量与各因素之间的关联程度, 可按一元线性回归分析处理, 如果没有因素能较好地解释因变量的变化状况, 再采用多元线性回归分析, 以分析各因素对水稻产量的显著性影响, 并进一步拟合出各因素对水稻产量的线性回归方程。

8.6.2 产品废品率的因素拟合

产品生产的质量往往受其中所用原料成份的影响, 特别是化学成分更会对产品的合格率产生影响。设某种产品生产过程中半成品的废品率与它含的一种化学成分有关, 经检验观测得到的一批数据如表 8.34 所示, 请确定产品废品率与化学成分之间的定量关系。(数据来源: 郝黎仁 等,《SPSS 实用统计分析》,中国水利水电出版社;参见数据文件: data8-7.sav。)

表 8.34 产品废品率与化学成分数据

序号	成分	废品率	序号	成分	废品率
1	34	1.3	9	40	0.44
2	36	1	10	41	0.56
3	37	0.73	11	42	0.3
4	38	0.9	12	43	0.42
5	39	0.81	13	43	0.35
6	39	0.7	14	45	0.4
7	39	0.6	15	47	0.41
8	40	0.5	16	48	0.6

案例分析: 产品中的成分应该控制在一定的比例范围内, 多了不行, 少了也不行。成分多少与废品率的关系不应该呈线性关系。同时从数据可以看出: 开始随着成分的逐渐增加, 废品率有逐渐降低的趋势, 当成分达到 42 时废品率最低, 但随着成分的进一步增加, 废品率又开始增加。可以看出, 成分与废品率之间呈非线性关系, 同时是一个自变量与因变量之间的关系, 故可采用曲线估计进行分析。

8.6.3 高管培训与表现预测

某著名总裁班的讲师想建立一个回归模型, 对参与培训的企业高管毕业后的长期表现情况进

行预测，观测到的数据如表 8.35 所示。自变量是高管的培训天数，因变量是高管毕业后的长期表现指数，指数越大，表现越好。（数据来源：杨维忠等，《SPSS 统计分析与行业应用案例详解》，清华大学出版社；参见数据文件：data8-8.sav。）

表 8.35 高管培训天数与长期表现指数

序号	培训天数	长期表现指数	序号	培训天数	长期表现指数
1	2	53	9	19	26
2	65	6	10	31	16
3	52	11	11	38	13
4	60	4	12	45	8
5	14	34	13	34	19
6	53	8	14	7	45
7	10	36	15	5	51
8	26	19			

案例分析：通过作散点图可知，自变量和因变量之间不成线性关系，可以采用曲线回归或非线性回归进行建模。进一步会发现两者之间成指数关系，在进行非线性回归时，可将模型设为 $y=\exp(a+b*x)$ 。并通过多种方法进行迭代，以求最佳拟合效果。

8.6.4 肾细胞癌转移的判断

肾细胞癌是否转移是临床治疗中选择治疗方案的重要依据。癌转移主要受以下几个因素的影响，即 x_1 ：确认时患者的年龄； x_2 ：肾细胞癌血管内皮生长因子（VEGF），其阳性表述由低到高共 3 个等级； x_3 ：肾细胞癌组织内微血管数（MVC）； x_4 ：肾癌细胞核组织学分级，由低到高共 4 级； x_5 ：肾细胞癌分期，由低到高共 4 期。用 y 代表肾细胞癌转移情况（有转移 $y=1$ ，无转移 $y=0$ ），数据如表 8.36 所示，根据肾细胞癌转移的影响因素对癌细胞是否转移进行预测判断。（数据来源：吕振通等，《SPSS 统计分析与应用》，机械工业出版社；参见数据文件：data8-9.sav。）

表 8.36 肾细胞癌转移数据资料

序号	x_1	x_2	x_3	x_4	x_5	y	序号	x_1	x_2	x_3	x_4	x_5	y
1	59	2	43.4	2	1	0	14	31	1	47.8	2	1	0
2	36	1	57.2	1	1	0	15	36	3	31.6	3	1	1
3	61	2	190	2	1	0	16	42	1	66.2	2	1	0
4	58	3	128	4	3	1	17	14	3	138.6	3	3	1
5	55	3	80	3	4	1	18	32	1	114	2	3	0
6	61	1	94.4	2	1	0	19	35	1	40.2	2	1	0
7	38	1	76	1	1	0	20	70	3	177.2	4	3	1
8	42	1	240	3	2	0	21	65	2	51.6	4	4	1
9	50	1	74	1	1	0	22	45	2	124	2	4	0
10	58	3	68.6	2	2	0	23	68	3	127.2	3	3	1
11	68	3	132.8	4	2	0	24	31	2	124.8	2	3	0
12	25	2	94.6	4	3	1	25	58	1	128	4	3	0
13	52	1	56	1	1	0	26	60	3	149.8	4	3	1

案例分析：癌细胞的转移情况分成两类：一类是没有转移，另一类是转移了。因变量是分类变量，不能用线性和非线性回归分析处理。这是一个典型的二元 Logistic 回归分析问题，根据 5 个影响因素进行预测判断。

8.7 思考与练习

1. 线性回归与非线性回归的关系是什么？
2. 在多元线性回归中，对回归方程作了检验后，为何还需对回归系数作检验？
3. 合金钢的强度 y 与钢材中碳的含量 x 有密切关系，为了冶炼出符合要求强度的钢，常常通过控制水中的碳含量来达到目的，因此需要了解 y 与 x 之间的关系，数据如表 8.37 所示，试对 x 和 y 进行一元线性回归分析。（参见数据文件：data8-10.sav。）

表 8.37 碳含量与钢强度数据

碳含量	0.03	0.04	0.05	0.07	0.09	0.1	0.12	0.15	0.17	0.2
钢强度	40.5	39.5	41	41.5	43	42	45	47.5	53	56

4. 一家大型商业银行设有 25 家分行，近年来其不良贷款额显著增加，其经营数据如表 8.38 所示（单位：亿元），能否将不良贷款与其他几个因素之间的关系用一定的数学关系式表达出来？如果能，用什么样的关系式表述它们之间的关系？能否用所建立的关系式预测出不良贷款？（数据来源：袁卫，《统计学（第二版）》，高等教育出版社；参见数据文件：data8-11.sav。）

表 8.38 某商业银行 25 家分行一年的主要业务数据

编号	不良贷款	各项贷款余额	本年累计应收贷款	贷款项目个数	本年固定资产投资额	编号	不良贷款	各项贷款余额	本年累计应收贷款	贷款项目个数	本年固定资产投资额
1	0.9	67.3	6.8	5	51.9	14	3.5	174.6	12.7	26	117.1
2	1.1	111.3	19.8	16	90.9	15	10.2	263.5	15.6	34	146.7
3	4.8	173.0	7.7	17	73.7	16	3.0	79.3	8.9	15	29.9
4	3.2	80.8	7.2	10	14.5	17	0.2	14.8	0.6	2	42.1
5	7.8	199.7	16.5	19	63.2	18	0.4	73.5	5.9	11	25.3
6	2.7	16.2	2.2	1	2.2	19	1.0	24.7	5.0	4	13.4
7	1.6	107.4	10.7	17	20.2	20	6.8	139.4	7.2	28	64.3
8	12.5	185.4	27.1	18	43.8	21	11.6	368.2	16.8	32	163.9
9	1.0	96.1	1.7	10	55.9	22	1.6	95.7	3.8	10	44.5
10	2.6	72.8	9.1	14	64.3	23	1.2	109.6	10.3	14	67.9
11	0.3	64.2	2.1	11	42.7	24	7.2	196.2	15.8	16	39.7
12	4.0	132.2	11.2	23	76.7	25	3.2	102.2	12.0	10	97.1
13	0.8	58.6	6.0	14	22.8						

5. 研究青春发育阶段的年龄与远视率的变化关系，测得数据如表 8.39 所示，请对年龄与远视率的关系进行曲线估计。（数据来源：袁卫，《统计学（第二版）》，高等教育出版社；参见数据文件：data8-12.sav。）

表 8.39 青春发育阶段年龄与远视率的变化关系

年龄 (x)	6	7	8	9	10	11	12	13	14	15	16	17	18
远视率 (y)	63.64	61.06	38.84	13.75	14.5	8.07	4.41	2.27	2.09	1.02	2.51	3.12	2.98

6. 棉花单株在不同时期的成铃数 (y) 与初花后天数 (x) 存在非线性的关系，假设这一非线性关系可用 Gompertz 模型表示： $y = a \cdot \exp(b \cdot \exp(c \cdot x))$ 。某一棉花品种 7 月 5 日至 9 月 3 日每隔 5 天的单株成铃数观测值如表 8.40 所示。试根据观测值拟合模型中的参数。（数据来源：朱军，《线性模型分析原理》，科学出版社；参见数据文件：data8-13.sav。）

表 8.40 棉花成铃数观测数据表

天数	5	10	15	20	25	30	35	40	45	50	55	60	65
成铃数	0.75	2	4	4.75	5.25	5.5	7.75	10.13	12.26	13.14	13.52	14.15	14.53

7. 在一次关于城镇居民上下班使用交通工具的调查中，因变量 $y = 1$ 表示居民主要乘坐公共汽车上下班； $y = 0$ 表示主要骑自行车上下班；其他因素主要包括年龄、月收入和性别（0 表示“女”，1 表示“男”）。数据如表 8.41 所示，试建立 y 与自变量间的 Logistic 回归模型。（数据来源：宋志刚 等，《SPSS 16 实用教程》，人民邮电出版社；参见数据文件：data8-14.sav。）

表 8.41 居民使用交通工具上下班情况

序号	年龄	月收入	性别	y	序号	年龄	月收入	性别	y
1	18	850	0	0	15	20	1000	1	0
2	21	1200	0	0	16	25	1200	1	0
3	23	850	0	1	17	27	1300	1	0
4	23	950	0	1	18	28	1500	1	0
5	28	1200	0	1	19	30	950	1	1
6	31	850	0	0	20	32	1000	1	0
7	36	1500	0	1	21	33	1800	1	0
8	42	1000	0	1	22	33	1000	1	0
9	46	950	0	1	23	38	1200	1	0
10	48	1200	0	0	24	41	1500	1	0
11	55	1800	0	1	25	45	1800	1	1
12	56	2100	0	1	26	48	1000	1	0
13	58	1800	0	1	27	52	1500	1	1
14	18	850	0	0	28	56	1800	1	1

第 9 章 聚类和判别分析

聚类分析（Cluster Analysis）和判别分析（Discriminant Analysis）都是研究事物分类的多元统计方法，两者紧密联系又有所区别。随着多元统计方法的快速发展和计算机的普遍应用，这两种方法在许多领域得到了大量应用，理论和软件也越来越成熟，已成为研究事物分类的最常用方法之一。SPSS 提供的“分析”菜单下的“分类”子菜单功能项用于解决这类问题，本章介绍聚类和判别分析的基本概念和统计原理，及常用的二阶聚类法、K-均值聚类法、系统聚类法、判别分析法等在 SPSS 中的实现过程。

9.1 聚类和判别分析简介

9.1.1 基本概念

人们认识某类事物时往往先对这类事物的各个对象进行分析，以便寻找同类事物的各种特征。如在国民经济领域，有时需要根据各省份的经济特点、产业结构、生产总值、人口数量、人均收入、消费特点等分成几个区域。比如分成经济发达地区、经济不发达地区、资源丰富地区、资源匮乏地区等。分成这样一些地区后，属于同一类地区的区域，国家可以采用类似的经济政策等。

在学生的学习生活中，我们也经常遇到这样的现象，有些学生关系比较密切，经常在一起，而与另外一部分同学则关系比较疏远。也就是说，学生根据他们自己的兴趣、爱好、学习成绩的好坏，会比较自然地形成一些固定的小群体。不同群体之间的学生兴趣爱好、家庭背景则存在比较明显的差异。

统计学研究这类问题的常用分类统计方法主要有聚类分析与判别分析。其中聚类分析是统计学中研究这种“物以类聚”问题的一种有效方法。聚类分析的基本思想是，认为研究的样本或指标（变量）之间存在着不同程度的相似性（亲疏关系）。于是根据一批样本的多个观测指标，具体找出一些能够度量样本或指标之间相似程度的统计量，以这些统计量为划分类型的依据，把一些相似程度较大的样本（或指标）聚合为一类，把另一些彼此之间相似程度较大的样本又聚合为另一类，关系密切的聚合到一个小的分类单位，关系疏远的聚合到一个大的分类单位，直到将所有的样本都聚合完毕，把不同的类型一一划分出来，形成一个由小到大的分类系统。最后再将整个分类系统画成一张谱系图，用它将所有样本间的亲疏关系表示出来。它是一种探索性的分析，在分类过程中，人们不必事先给出一个分类标准，聚类分析能够从样本数据出发，自动进行分类。根据所使用的方法不同，常常会得到不同的结论。

根据分类对象的不同，聚类分析可分为对样本的聚类和对变量的聚类两种。

- 样本聚类：也称为 Q 型聚类，是对观测量（Case）进行聚类，本书主要研究如何解决样本聚类的方法和应用。
- 变量聚类：也称为 R 聚类，能够找出彼此独立且有代表性的自变量，而又不丢失大部分信息，主要对研究对象的观测变量进行聚类，将具有相同特性的变量作为一类。在生产活动中不乏变量聚类的实例，如衣服号码（身高、胸围、裤长、腰围）、鞋的尺码；在儿童的生长发育研究中，把以形态学为主的指标归为一类，以机能为主的指标归为另一类。

判别分析是判别样本所属类型的一种统计方法。与聚类分析一样，判别分析也用于解决分类问题，不同之处在于，判别分析在已知研究对象分为若干类型（或组别）并已取得各种类型的一批已知样本的观测数据的基础上，根据某些准则建立判别式，然后对未知类型的样本进行判别分析。

例如，银行为了对贷款进行管理，需要预测哪些类型的客户可能不会按时归还贷款。已知过去几年中 1000 个客户的贷款归还信誉度，据此可以将客户分成两组：可靠客户和不可靠客户。再通过收集客户的一些资料，如年龄、工资收入、教育程度、存款等，将这些资料作为自变量，通过判别分析，建立判别函数。如果有新客户提交贷款申请，就可以利用创建好的判别函数，对新客户进行分析，如果新客户属于可靠类就放款给他，否则就不放款。

进行判别分析的方法很多，按照判别准则可以分为距离判别、Bayes 判别和 Fisher 判别等。

9.1.2 样本间亲疏关系的度量

1. 连续变量的样本间距离常用度量

样本若有 k 个变量，则可以将样本看成 k 维空间中的一个点，样本和样本之间的距离就是 k 维空间中点与点之间的距离，这反映了样本之间的亲疏程度。聚类时，距离近的样本属于同一类，距离远的样本属于不同类。

主要方法有欧氏距离（Euclidean Distance）、欧氏平方距离（Squared Euclidean Distance）、切比雪夫距离（Chebychev Distance）、明可夫斯基距离（Minkowski Distance）、用户自定义距离（Customize Distance）等，这些方法的模型参见 7.4 节表 7.9。

除上面的方法外，还有其他几种方法，如 Pearson 相关系数式（7.1）、夹角余弦（Cosine），夹角余弦的模型如式（9.1）。

$$c_{ij} = \cos \alpha_{ij} = \frac{\sum_{k=1}^n x_{ki} x_{kj}}{\sqrt{\sum_{k=1}^n x_{ki}^2 \sum_{k=1}^n x_{kj}^2}}$$

(9.1)

它是两变量预测值 (x_{1i}, \cdots, x_{ni}) 与 (x_{1j}, \cdots, x_{nj}) 之间夹角 α_{ij} 的余弦函数，从数据矩阵来看，就是数据矩阵第 i 列和第 j 列向量的夹角余弦。

2. 顺序变量的样本间距离常用度量

常用的有 χ^2 统计量（Chi-square measure）和 ϕ^2 统计量（Phi-square measure），具体计算公式参见 7.4 节表 7.10。

☆说明☆

- (1) 聚类分析的目的是找到样本中数据的特点，因此应注意所选择的变量是否已经能够反映所要聚类样本的主要特点。
- (2) 聚类分析时应注意所选择的变量是否存在数量级上的差别。如果一个样本包含不同数量级的变量，则应先对变量进行标准化处理，而后再进行聚类。
- (3) 变量间的关系度量模型与样本间相类似，只不过一个用矩阵的行进行计算，另一个用矩阵的列进行计算。

9.2 二阶聚类

9.2.1 基本概念及统计原理

1. 基本概念

二阶聚类 (Twostep Cluster) 是一个探索性的分析工具, 为揭示自然的分类或分组而设计, 是数据集内部的而不是外观上的分类, 是一种新型的分层聚类算法 (Hierarchical Algorithms)。目前主要应用在数据挖掘和多元数据统计的交叉领域——模式分类中, 其算法适合任何尺度的变量。二阶聚类分析主要利用距离度量, 假设聚类模型的变量均为自变量, 即假设连续型变量为正态分布, 分类变量是多项式。

该过程主要有以下几个特点: (1) 分类变量和连续变量均可以参与二阶聚类分析; (2) 该过程可以自动确定分类数; (3) 可以高效率地分析大数据集; (4) 用户可以自己定制用于运算的内存容量。

2. 统计原理

二阶聚类的功能非常强大, 而原理又较为复杂。在聚类过程中除了使用传统的欧氏距离外, 为了处理分类变量和连续变量, 它用似然距离测度, 并要求模型中的变量是独立的。分类变量呈多项式分布, 连续变量呈正态分布。

使用两个变量的相关过程 (Bivariate Correlations) 去检验两个连续变量之间的独立性, 使用交叉表 (Crosstabs) 过程检验两个分类变量之间的独立性, 使用均值比较 (Means) 过程检验连续变量与分类变量的独立性, 用探索分析过程检验连续变量的正态性, 使用卡方检验 (Chi-Square Test) 过程检验分类变量是否呈多项式分布。

3. 分析步骤

二阶聚类分成两个步骤完成。

第1步 构建聚类特征树。

对每个观测变量考察一遍, 确定类中心。根据相近者为同一类的原则, 计算距离并把与类中心距离最小的观测量分到相应的类中去, 这个过程称为构建一个分类的特征树 (CF)。开始, 它把一个观测量放在树的叶节点根部, 该节点含有该观测量的变量信息; 然后, 使用距离测度作为相似性测度判据, 每个后续的观测量根据它在已经存在的节点的相似性归到某类中去。如果相似则将该观测量加在一个已经存在的节点上, 形成该节点的树叶; 如果不相似, 就形成一个新的节点。

第2步 对聚类特征树的节点进行分组。

为确定最好的类数, 对每一个聚类结果使用 Akaik 判据 (AIC) 或贝叶斯判据 (BIC) 作为标准进行比较, 得出最后的聚类结果。

9.2.2 二阶聚类 SPSS 实例分析

【例 9-1】某机构为了调查学生性别和所学专业与毕业后初始工资的情况, 调查抽取了 60 名学生的数据, 如表 9.1 所示 (其中“性别”1 代表男性, 0 代表女性; “学科”1 代表农学, 2 代表建筑, 3 代表地质, 4 代表商务, 5 代表林学, 6 代表教育, 7 代表工程, 8 代表艺术), 试根据样本指标进行聚类分析。(参见数据文件: data9-1.sav。)

表 9.1 学生信息表

序号	性别	学科	工资	序号	性别	学科	工资	序号	性别	学科	工资
1	1	7	28900	21	1	7	29000	41	1	7	28200
2	1	7	28000	22	1	7	32000	42	1	1	15000
3	1	1	27500	23	1	7	33500	43	0	1	27000
4	1	7	30300	24	1	7	27000	44	1	4	30000
5	1	1	18000	25	0	1	29000	45	0	4	18800
6	0	7	31700	26	1	4	19000	46	0	4	21500
7	1	3	26000	27	0	8	20900	47	1	3	23000
8	1	7	25000	28	0	1	29000	48	0	7	25500
9	0	1	20000	29	0	1	35300	49	1	4	25000
10	1	1	18000	30	0	1	24200	50	0	1	13500
11	1	4	23000	31	1	3	41000	51	1	4	23600
12	1	4	27600	32	1	7	36300	52	0	4	19000
13	1	7	32700	33	0	6	23000	53	1	7	30600
14	0	1	21500	34	1	4	25000	54	1	1	27500
15	1	1	25000	35	1	4	18200	55	0	1	26300
16	0	4	18000	36	1	7	25400	56	0	4	30000
17	1	7	38400	37	1	1	24000	57	1	4	24000
18	0	1	26500	38	0	1	20000	58	1	7	28000
19	0	1	26500	39	0	4	22000	59	1	7	27100
20	0	1	31000	40	0	7	32000	60	1	7	26400

第 1 步 分析。

由于自变量中不仅有连续属性，也有分类变量，故采用二阶聚类进行分析。

第 2 步 数据组织。

按表所示定义变量，输入数据并保存。

第 3 步 二阶聚类设置。

(1) 按“分析→分类→二阶聚类”顺序打开“二阶聚类分析”对话框，并按图 9-1 所示进行设置。

该对话框主要由以下几部分组成。

- ① “分类变量”框：用于放置离散变量，也可以放入连续变量，这时系统将把连续变量当作离散变量来处理。
- ② “连续变量”框：用于放置连续变量，离散变量无法移入其中。
- ③ “距离测量”选项组：用于距离的测量方法。其中包括两个选项：对数似然值和欧氏 (Euclidean) 距离，前者为系统默认值。当没有选入离散变量时，读者可以任意选择这两种方法中的一种，不过如果选择欧氏的话，相当于使用传统聚类方法进行聚类；当有离散变量选入时，欧氏选项将无法使用，只能使用对数似然值。
- ④ “连续变量计数”框组：用于显示选入的连续变量数目及状态。
- ⑤ “聚类数目”框组：包含两个选项，第一项是指由系统自动决定的分类数目，并在下面的“最大值”栏内输入一个数值来限制分类的最大数目，此选项为系统默认选项；第二项是指由客户自己确定分类数目，在下面的“数量”栏内输入这个指定值。
- ⑥ “聚类准则”选项： SPSS 提供了两个准则，分别是 BIC (Bayesian Information Criterion)

准则和 AIC (Akaike Information Criterion) 准则, 这两个指标越小, 聚类效果越好。系统会根据 BIC 和 AIC 的大小, 以及类间最短距离的变化情况来确定最优的聚类类别数。

(2) “选项”对话框设置: 单击“选项(O)…”按钮, 弹出此子对话框, 如图 9-2 所示。



图 9-1 “二阶聚类分析”对话框

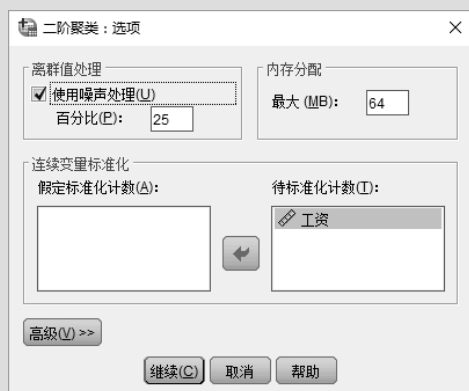


图 9-2 “二阶聚类: 选项”对话框

现对各项的意义解释如下。

① “离群值处理”框: 用于指定在聚类过程中当产生聚类特征树时奇异值的处理方式, 当选择“使用噪声处理”选项时, 需要在“百分比”文本框中输入百分比数值。

② “内存分配”框: 用于指定聚类计算时的最大内存 (MB), 系统默认为 64 MB, 一般使用系统默认值。

③ “连续变量标准化”框: 从左侧的“假定标准化计数”列表框中指定需要标准化的连续型变量, 将其移动到右侧“待标准化计数”列表框。因为聚类算法中要求连续型变量必须是标准化变量, 所以应该将所有连续型变量都移至右侧列表框, 以便在聚类计算之前标准化所有的连续型变量, 这样可以减少计算量, 提高聚类效率。

④ “高级”按钮: 主要用于对前面提到的聚类特征树的选项进行设置, 一般使用系统默认值。

(3) “输出”对话框设置: 单击“输出(U)…”按钮, 弹出此子对话框, 如图 9-3 所示。

① “输出”选项: 选中“图表和表”后, 会将聚类分析的概要表显示出来。

② “工作数据文件”选项: 选中“创建聚类成员变量”用于在文件中创建一个新变量, 保存各个观测量的所属类别。

③ “XML 文件”选项: 选择输出聚类的最终模型或聚类特征树到指定位置。

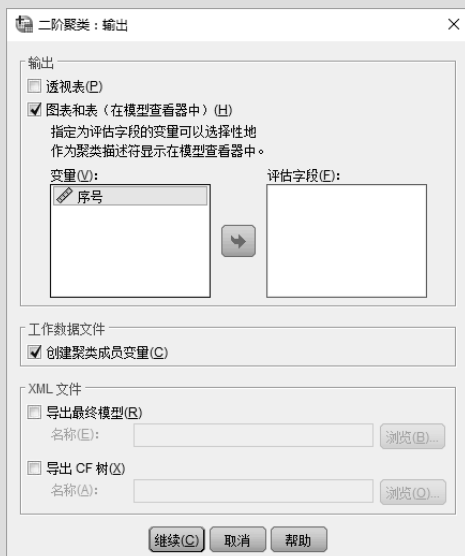


图 9-3 “二阶聚类: 输出”对话框

以上各项设置完成后提交系统运行。

第 4 步 主要结果及分析。

运行结果如图 9-4~图 9-6 所示，分别解释如下：

(1)图 9-4 是二阶聚类的模型概要和聚类质量情况。从中可以看出，此算法采用的是两步(二阶)聚类，共输入 3 个变量，将所有个案聚成 3 类。聚类的平均轮廓值为 0.6 (其范围值为-1.0~1.0，值越大越好)，说明聚类质量较好。

(2)图 9-5 是在 SPSS 输出框中双击图 9-4 所显示的结果，可以看出各类所占的比例情况。

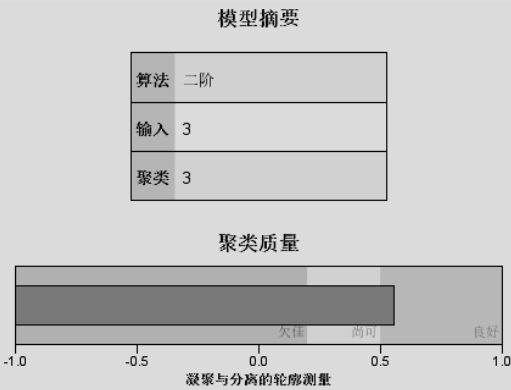


图 9-4 模型摘要与聚类质量

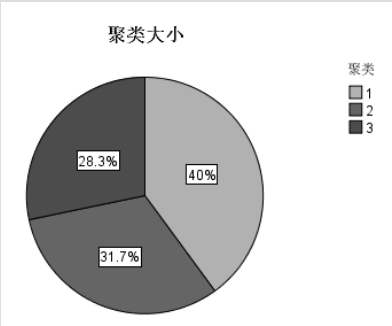


图 9-5 各类比例情况

(3)图 9-6 显示了数据文件中聚类后各个案所属的分类号情况，最后一列表示各条数据二阶聚类的类别号。

序号	性别	学科	工资	TSC_8162
1	1	7	28900	3
2	1	7	28000	3
3	1	1	27500	2
4	1	7	30300	3
5	1	1	18000	2
6	0	7	31700	1
7	1	3	26000	2
8	1	7	25000	3
9	0	1	20000	1
10	1	1	18000	2
11	1	4	23000	2
12	1	4	27600	2
13	1	7	32700	3
14	0	1	21500	1
15	1	1	25000	2

图 9-6 数据表中保存个案的类别号

9.3 K-均值聚类

9.3.1 基本概念及统计原理

1. 基本概念

K-均值聚类是由用户指定类别数的大样本资料的逐步聚类分析方法。它先对数据进行初始分

类，然后逐步调整，得到最终分类数。当要聚成的类数已知时，使用 K-均值聚类的处理速度快，占用的计算机内存少。

2. 统计原理

K-均值聚类执行 Quick Cluster 命令，首先要选择用于聚类分析的变量和类数。参与聚类分析的变量必须是数值型。为了清楚地表明各观测量最后聚到哪一类，还应该指定一个表明观测量特征的变量作为标识变量，例如编号、姓名等。聚类数必须大于等于 2。

如果选择了 n 个数值型变量参与聚类分析，最后要求聚类数为 k ，那么可以由系统首先选择 k 个观测量（也可以由用户指定）作为聚类目标， n 个变量组成 n 维空间。每个观测量在 n 维空间中是一个点。 k 个事先选定的观测量就是 k 个聚类中心点，也称为初始类中心。按照与这几个类中心的距离（使用的是欧氏距离（Euclidean Distance））最小原则将观测量分派到各类中心所在的类中去，构成第一次迭代形成的 k 类，根据组成每一类的观测量，计算各变量均值。每一类中的 n 个均值在 n 维空间中又形成 k 个点，这就是第二次迭代的类中心。按照这种方法依次迭代下去，直到达到指定的迭代次数或达到中止迭代的判据要求时，迭代停止，聚类过程结束。

3. 分析步骤

第 1 步 指定聚类数目 k 。

由用户指定。

第 2 步 确定 k 个初始类中心。

在指定了聚类数目 k 后，还需要指定这 k 个初始类的中心点。中心点的指定方式有两种：第一，用户指定方式，用户应事先准备好一个存有 k 个样本的 SPSS 数据文件，这 k 个样本将作为 k 个类的初始类中心点，这就需要用户根据实际问题的分析需要和以往经验指定相对合理的初始类中心；第二，系统指定方式，SPSS 会根据样本数据的具体情况选择 k 个有一定代表性的样本作为初始中心点。

第 3 步 根据距离最近原则进行聚类。

依次计算每个样本数据点到 k 个类中心点的欧氏距离，并按照 k 个类中心点距离最短的原则将所有样本分派，聚成 k 个类。

第 4 步 重新确定 k 个类中心。

SPSS 计算每个类中各变量的均值，并以均值点作为新的类中心点。

第 5 步 迭代计算。

重复第 3 步和第 4 步，直到达到指定的迭代次数或终止迭代的判据要求为止。

9.3.2 K-均值聚类 SPSS 实例分析

【例 9-2】 测量 12 名大学生对“高等数学”课程的心理状况和学习效果，主要包括四个因素：学习动机、学习态度、自我感觉、学习效果，具体数据如表 9.2 所示。试将该 12 名学生分成 3 类以分析不同心理状况下学生的学习效果。（参见数据文件：data9-2.sav。）

表 9.2 12 名学生学习课程的心理变化数据

编号	学习动机	学习态度	自我感觉	学习效果
1	40	80	54	44
2	37	73	56	46
3	43	70	75	58
4	50	77	85	77

续表				
编号	学习动机	学习态度	自我感觉	学习效果
6	67	70	84	69
7	77	37	57	100
8	80	37	73	82
9	83	40	76	96
10	87	43	75	91
11	60	57	70	85
12	70	50	69	90

第 1 步 分析。

由于已知分成 3 类，故可采用 K-均值聚类法。

第 2 步 数据组织。

按表 9.2 所示组织数据，将“编号”变量的数据类型设为字符型（作为标识变量）。

第 3 步 K-均值聚类设置。

（1）按“分析→分类→K-均值聚类”顺序打开“K-均值聚类分析”对话框，将“学习动机”、“学习态度”、“自我感觉”、“学习效果”四个变量选入“变量”列表框。将“编号”变量移入“个案标注依据”框中；将“聚类数”设为 3。具体如图 9-7 所示。



图 9-7 “K-均值聚类分析”对话框

对其中的几个选项（组）解释如下。

- ① “变量”框：用于放置进行 K-均值聚类的变量。
- ② “个案标注依据”框：用于标志各观测值的所属类的变量，相当于观测量记录号的作用。
- ③ “聚类数”框：用于设置聚类数目，默认值为 2。
- ④ “方法”选项：用于选择聚类方法。系统默认选项是“迭代与分类”，该选项是指在迭代过程中不断地更新聚类中心；“仅分类”是指迭代过程中聚类中心一直不变。
- ⑤ “聚类中心”选项组：用于设置初始聚类中心和最终聚类中心的存取。其中“读取初始聚类中心”表示从文件或数据集中读取初设的聚类中心，“写入最终聚类中心”表示将最终聚类中心保存到指定的文件或数据集中。

(2) “迭代”对话框设置: 单击“迭代(I)…”按钮, 弹出此子对话框, 并按图 9-8 所示设置。

① “最大迭代次数”框: 栏内输入迭代次数的上限, 系统默认为 10。

② “收敛条件”框: 栏内输入一个不超过 1 的正数, 其默认值为 0。若输入数值为 0.02, 表示两迭代计算的最小类中心的变化距离小于初始类中心距离的 2% 时迭代停止。

③ “使用运行平均值”选项: 选择此项表示在迭代过程中每分配一个观测量到某类后就立刻计算新的聚类中心, 不选此项表示当所有观测量分配完以后再计算各聚类中心。

(3) “保存”对话框设置: 单击“保存(S)…”按钮, 弹出此子对话框, 并按图 9-9 所示设置。

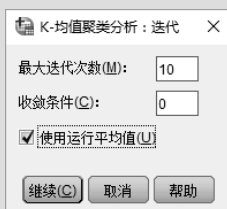


图 9-8 “K-均值聚类分析: 迭代”对话框

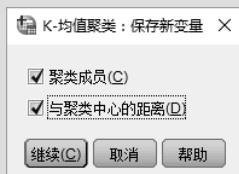


图 9-9 “K-均值聚类: 保存新变量”对话框

① “聚类成员”选项: 选择该项后, 数据文件中将新建一个名为“QCL_1”的变量, 其值为各观测变量的类别。

② “与聚类中心的距离”选项: 若选择此项, 工作文件中将建立一个名为“QCL_2”的变量, 其值为各观测变量与所属类的类中心之间的欧氏距离。

(4) “选项”对话框设置: 单击“选项(O)…”按钮, 弹出此子对话框, 并按图 9-10 所示设置。

① “统计”选项组: 用于指定输出统计量值, 包括以下几类。

➤ “初始聚类中心”选项: 输出初始聚类中心, 为系统默认选项。

➤ “ANOVA 表”选项: 方差分析表选项, 输出方差分析表。在聚类过程中, 可能引入了无关变量, 这样会降低聚类的效果。可见, 使用方差分析表来分析变量在类间的差异, 若发现差异很小的变量, 就可以将它从“变量框”中去除。(在聚类分析时, 并不是变量越多越好, 差异很小的变量可能会影响分类的准确性。)

➤ “每个个案的聚类信息”选项: 每个观测量的聚类信息选项, 显示每个观测量最终被聚入的类别、各个观测量与最终聚类中心的欧氏距离, 以及最终各类之间的欧氏距离。

② “缺失值”选项组: 用于指定缺失值的处理方式。第一个选项为系统默认选项, 指聚类分析中凡是有缺失值的观测量均剔除。第二个选项指聚类变量中只有有缺失值的观测量才予以剔除。

以上各项设置完后提交系统运行。

第4步 主要结果及分析。

运行结果如表 9.3~表 9.6 和图 9-11 所示, 分别解释如下。

(1) 表 9.3 是初始聚类中心表, 由于没有指定初始聚类中心, 故列出了由系统指定的聚类中心。与原数据比较, 可见它们分别是第 1 号、第 6 号和第 7 号个案。

(2) 表 9.4 是迭代历史表。由表可知, 第一次迭代后, 3 个类的中心点分别变化了 8.193, 9.889 和 13.472。一共进行了 10 次迭代, 达到聚类结果的要求 (达到最大迭代次数), 聚类分析结束。

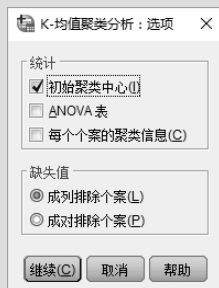


图 9-10 “K-均值聚类分析: 选项”对话框

表 9.3 初始聚类中心

	聚类		
	1	2	3
学习动机	40	67	77
学习态度	80	70	37
自我感觉	54	84	57
学习效果	44	69	100

(3) 表 9.5 是聚类结果的类中心表。如第 1 类的学习动机值为 39，学习态度值为 77，自我感觉值为 55，学习效果值为 45。

(4) 表 9.6 是每类中包含的样本数情况。可看出第 1，2，3 类中分别含有 2，4，6 个样本。

(5) 查看数据文件，可看到多出两个变量，分别表示每个个案的具体分类归属和与类中心的距离，具体如图 9-11 所示。

表 9.5 最终聚类中心

	聚类		
	1	2	3
学习动机	39	52	76
学习态度	77	76	44
自我感觉	55	83	70
学习效果	45	67	91

表 9.4 迭代历史表^a

迭代	聚类中心中的变动		
	1	2	3
1	8.193	9.889	13.472
2	3.909	7.631	4.701
3	1.303	1.526	.672
4	.434	.305	.096
5	.145	.061	.014
6	.048	.012	.002
7	.016	.002	.000
8	.005	.000	3.996E-5
9	.002	9.768E-5	5.709E-6
10	.001	1.954E-5	8.155E-7

a. 由于已达到最大迭代执行次数，因此迭代已停止。迭代未能收敛。任何中心的最大绝对坐标变动为 .000。当前迭代为 10。初始中心之间的最小距离为 48.518。

表 9.6 每类中的个案数

聚类	1	2.000
	2	4.000
	3	6.000
有效	12.000	
缺失	.000	

编号	学习动机	学习态度	自我感觉	学习效果	QCL_1	QCL_2
1	40	80	54	44	1	4.062
2	37	73	56	46	1	4.062
3	43	70	75	58	2	16.037
4	50	77	85	77	2	10.592
5	47	87	89	63	2	13.809
6	67	70	84	69	2	16.559
7	77	37	57	100	3	17.487
8	80	37	73	82	3	12.158
9	83	40	76	96	3	11.276
10	87	43	75	91	3	11.978
11	60	57	70	85	3	21.505
12	70	50	69	90	3	8.687

图 9-11 保存到数据文件中的变量

9.4 系统聚类

9.4.1 基本概念及统计原理

1. 基本概念

系统聚类是效果最好且经常使用的方法之一，国内外对它进行了深入的研究，系统聚类在聚

类过程中是按一定层次进行的。具体分成两种，分别是 Q 型聚类和 R 型聚类，Q 型聚类是对样本（个案）进行的分类，它将具有共同特点的个案聚集在一起，以便对不同类的样本进行分析；R 型聚类是对变量进行的聚类，它使具有共同特征的变量聚在一起，以便对不同类的变量进行分析。

2. 统计原理

系统聚类是根据个案或变量之间的亲疏程度，将最相似的对象聚集在一起。根据系统聚类过程的不同，又分为凝聚法和分解法两种。凝聚法的原理是将参与聚类的每个个案（或变量）视为一类，根据两类之间的距离或相似性，逐步合并直到合并为一个大类为止；分解法的原理是将所有个案（或变量）都视为一类，然后根据距离和相似性逐层分解，直到参与聚类的每个个案（或变量）自成一类为止。实际上以上两种方法是方向相反的两种聚类过程。

在系统聚类中，度量数据之间的亲疏程度是极为关键的。在衡量样本与样本之间的距离时，一般使用的距离有欧氏距离、欧氏平方距离、切比雪夫距离、Block 距离、明可夫斯基距离、夹角余弦等，具体模型如本章前面所述。

衡量样本数据与小类、小类与小类之间亲疏程度的度量方法主要有以下 7 种。

(1) 最短距离法 (Nearest Neighbor)：以当前某个样本与已形成小类中各样本距离的最小值作为当前样本与该小类之间的距离。

(2) 最长距离法 (Furthest Neighbor)：以当前某个样本与已形成小类中各样本距离的最大值作为当前样本与该小类之间的距离。

(3) 类间平均链锁法 (Between-groups Linkage)：两小类之间的距离为两个小类所有样本间的平均距离。

(4) 类内平均链锁法 (Within-groups Linkage)：与小类间平均链锁法类似，这里的平均距离是对所有样本对的距离求平均值，包括小类之间的样本对、小类内的样本对。

(5) 重心法 (Centriod Clustering)：将两小类间的距离定义成两小类重心间的距离。每一小类的重心就是该类中所有样本在各个变量上的均值代表点。

(6) 中间距离法 (Median Clustering)：以两类变量均值之间的距离作为类与类之间的距离。

(7) 离差平方和 (Ward's Method)：在聚类过程中，使小类内各个样本的欧氏距离总平方和增加最小的两小类合并成一类。

9.4.2 系统聚类 SPSS 实例分析

【例 9-3】 已知 29 例儿童的血中血红蛋白 (Hemoglobin, g)、钙 (Ca, ug)、镁 (Mg, ug)、铁 (Fe, ug)、锰 (Mn, ug)、铜 (Cu, ug) 的含量如表 9.7 所示，试对数据进行变量聚类分析。（参见数据文件：data9-3.sav。）

表 9.7 儿童血液中的微量元素含量

序号	Ca	Mg	Fe	Mn	Cu	Hemogl	序号	Ca	Mg	Fe	Mn	Cu	Hemogl
1	54.89	30.86	448.7	0.012	1.01	13.5	9	60.35	38.2	394.4	0.001	1.14	12
2	72.49	42.61	467.3	0.008	1.64	13	10	54.04	34.23	405.6	0.008	1.3	11.75
3	53.81	52.86	425.61	0.004	1.22	13.75	11	61.23	37.35	446	0.022	1.38	11.5
4	64.74	39.18	469.8	0.005	1.22	14	12	60.17	33.67	383.2	0.001	0.914	11.25
5	58.8	37.67	456.55	0.012	1.01	14.25	13	69.69	40.01	416.7	0.012	1.35	11
6	43.67	26.18	395.78	0.001	0.594	12.75	14	73.89	32.94	312.5	0.064	1.15	7.25
7	54.89	30.86	448.7	0.012	1.01	12.5	15	47.31	28.55	294.7	0.005	0.838	7
8	86.12	43.79	440.13	0.017	1.77	12.25	16	72.28	40.12	430.8	0	1.2	10.75

							续表						
序号	Ca	Mg	Fe	Mn	Cu	Hemogl	序号	Ca	Mg	Fe	Mn	Cu	Hemogl
17	55.13	33.02	445.8	0.012	0.918	10.5	24	61.02	29.27	258.94	0.016	1.19	8.75
18	70.08	36.81	409.8	0.012	1.19	10.25	25	53.68	28.79	292.8	0.048	1.32	8.5
19	63.05	35.07	384.1	0	0.853	10	26	50.22	29.17	292.6	0.006	1.04	8.25
20	48.75	30.53	342.9	0.018	0.924	9.75	27	65.34	29.99	312.8	0.006	1.03	8
21	52.28	27.14	326.29	0.004	0.817	9.5	28	56.39	29.29	283	0.016	1.35	7.8
22	52.21	36.18	388.54	0.024	1.02	9.25	29	66.12	31.93	344.2	0	0.689	7.5
23	49.71	25.43	331.1	0.012	0.897	9							

第 1 步 分析。

根据题目要求，需进行变量聚类分析（即 R 型聚类），故采用系统聚类分析中的 R 型聚类进行处理。

第 2 步 数据组织。

定义 7 个变量：“序号”、“ca”（钙）、“mg”（镁）、“fe”（铁）、“mn”（锰）、“cu”（铜）和“hemogl”（血红蛋白），其中“序号”为字符型，其余变量为数值型，输入数据并保存。

第 3 步 进行按变量聚类的设置。

（1）按“分析→分类→系统聚类”打开“系统聚类分析”对话框，将“ca”（钙）、“mg”（镁）、“fe”（铁）、“mn”（锰）、“cu”（铜）和“hemogl”（血红蛋白）几个变量选入“变量”列表框。设置按“变量”分类，并选择输出“统计”（量）和“图”，以激活“统计(S)…”和“图(T)…”两个按钮。具体如图 9-12 所示。

对该对话框的各项解释如下。

- ①“变量”框：选择需要用于聚类分析的变量。
- ②“个案标注依据”框：用于放置标记变量，相当于观测量记录号的作用，变量类型只能是字符类型。
- ③“聚类”选项组：用于选择聚类类型，“个案”是按观测量的样本进行聚类，即 Q 型聚类。而“变量”是按变量进行聚类，即 R 型聚类。
- ④“显示”选项组：选择显示内容，选中“统计”激活“统计(S)…”按钮，并可进行相应的统计量输出设置，类似地，选中“图”则激活“图(T)…”按钮，并可进行相应的图形输出设置。

（2）“统计”对话框设置：单击“统计(S)…”按钮，弹出此子对话框，设置如图 9-13 所示。



图 9-12 “系统聚类分析”对话框



图 9-13 “系统聚类分析：统计”对话框

现对其中各项解释如下。

① “集中计划”选项：系统默认选项，输出一张概述聚类进程的表格，反映聚类过程中每一步样本或变量的合并情况。

② “近似值矩阵”选项：显示各项间的距离矩阵。

③ “聚类成员”选项组：包含以下三项。

➤ “无”选项：不输出样品隶属类表，为系统默认选项；

➤ “单个解”选项：选择此项并在下边的“聚类数”框中指定表示分类数的一个大于1的整数，则输出各样本或变量的隶属表；

➤ “解的范围”选项：指定两个分类数 $m < n$ ，输出分类数从 m 到 n 的各种分类的样本隶属表。

(3) “图”对话框设置：单击“图(T)...”按钮，弹出此子对话框，设置如图9-14所示，对其中的各选项解释如下。

① “谱系图”选项：选择此项将输出反映聚类结果的龙骨图(树形图)。

② “冰柱图”选项组：包含以下四项。

➤ “全部聚类”选项：显示全部聚类结果的冰柱图。

➤ “指定范围内的聚类”选项：限制聚类解范围，在下面的“开始聚类”、“停止聚类”和“依据”3个小框中分别输入3个正整数 m, n, k ($m \leq n, k \leq n$)，表示从最小聚类解 m 开始，以增量 k 为步长，到最大聚类解 n 为止。

➤ “无”选项：不输出冰柱图。

➤ “方向”选项组：以“垂直”或“水平”形式输出冰柱图。

(4) “方法”对话框设置：单击“方法(M)...”按钮，弹出此子对话框，设置如图9-15所示，对其中的各选项解释如下。



图 9-14 “系统聚类分析：图”对话框

图 9-15 “系统聚类分析：方法”对话框

① “聚类方法”下拉列表：可以选择“组间联接”、“组内联接”、“最近邻元素”、“最远邻元素”、“质心聚类”、“中位数聚类”和“瓦尔德法”7种方法中的一种。

☆说明☆

- ◆ 由于不同聚类方法所使用的聚类模型不一样，选用不同的聚类方法，所得到的聚类结果可能会有很大区别。

- ② “测量”选项组：用于选择距离测度方法下面的 3 个选项。
- “区间”选项：为连续型变量提供距离算法，其中默认为“欧氏距离”，其他还有“平方欧氏距离”、“余弦”、“皮尔逊相关性”、“切比雪夫”、“块”、“明可夫斯基”和“定制”7 种方法，这些方法的模型参见 7.4 节。
 - “二元”选项：为二元变量提供的二值数据的不相似性测度，其中默认为平方欧氏距离。
 - “计数”选项：为分类变量提供了卡方测量和 Phi 平方测量两种距离测量方法、其计算公式如表 7.10 所示。
- ③ “转换值”选项组：用于选择数据标准化方法。
- SPSS 默认不进行标准化处理，如果需要，也可选择下拉菜单的标准化方法，包括正态标准化（Z 得分）、全距从-1 到 1、全距从 0 到 1、1 的最大量、均值为 1、标准差为 1 等几种方法。
 - 如果选择了一种标准化处理方法，则需要指定标准化处理针对的是“按变量”还是“按个案”。
- ④ “转换测量”选项组：用于选择转换方法。系统提供了 3 种方法，包括“绝对值”法、“更改符号”法、“重新标度到 0-1 范围”法。

（5）“保存”子对话框设置：用于保存新变量，只有对观测变量进行聚类时，此项才被激活。其结果不是以表格的形式输出，而是保存到数据窗口中，其“聚类成员”的保存解释与图 9-13 下半部分的解释一致。

第 4 步 主要结果及分析。

主要结果如表 9.8、图 9-16 和图 9-17 所示，分别解释如下。

（1）表 9.8 是聚类顺序表，第 1 步是第 4 个变量和第 5 个变量最先进行了聚类，变量间的距离系数为 6.028，这个聚类的结果将在后面的第 2 步聚类中用到；第 2 步是经过第 1 步聚类后的变量 4 和变量 5 与变量 6 进行聚类，变量间的距离系数为 54.938，这个聚类的结果将在第 4 步中用到。以此类推，这 6 个变量经过 5 步聚类最终聚成一个大类。

表 9.8 聚类顺序表

阶段	组合聚类		系数	首次出现聚类的阶段		下一个阶段
	聚类 1	聚类 2		聚类 1	聚类 2	
1	4	5	6.028	0	0	2
2	4	6	54.938	1	0	4
3	1	2	144.078	0	0	4
4	1	4	235.530	3	2	5
5	1	3	1966.192	4	0	0

（2）图 9-16 是系统聚类的冰柱图，由于聚类过程像冰柱的形状而得名。图的纵坐标表示聚类的数目，我们从图的最下方看起，从 5 类，逐渐到 4 类、3 类、2 类，最后聚成一个大类。首先是“铜”和“锰”聚成一类，其余每个变量各为一类。第 2 步再将“血红蛋白”聚到“铜”和“锰”一类中，原先的 6 个变量就变成了 4 类。以此类推，经过 5 步聚类，最后将所有变量聚成了一个 1 类。

（3）图 9-17 是系统聚类的谱系图，从中可看出，第 1 步将“cu（铜）”和“mn（锰）”聚成一类，第 2 步将“hemogl（血红蛋白）”聚到“cu（铜）”和“mn（锰）”类中，第 3 步将“ca（钙）”和“mg（镁）”聚成一类。以此类推，最后聚成一个大类。这与表 9.8 所示聚类顺序表和图 9-16 所示聚类冰柱图的分析结果是一致的。

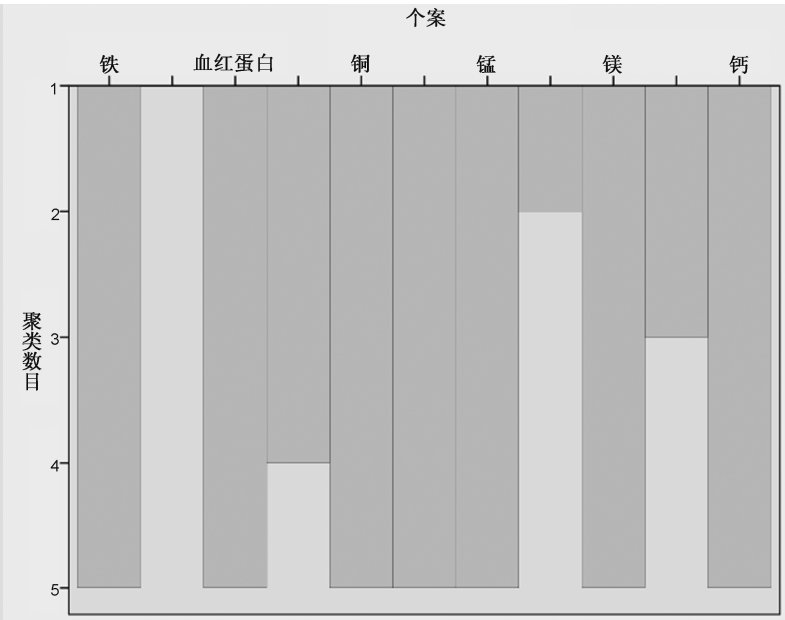


图 9-16 聚类冰柱图

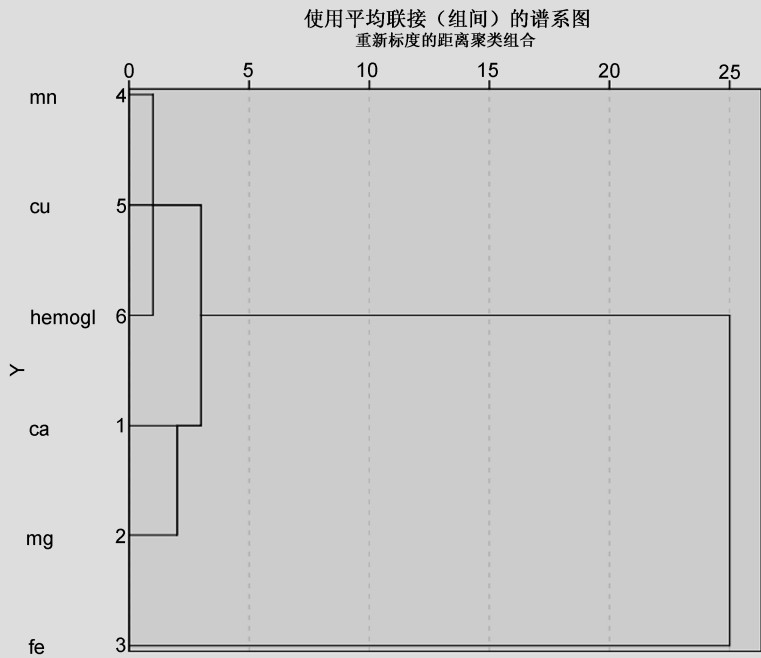


图 9-17 系统聚类的谱系图

☆说明☆

◆ 例 9-3 是一个 R 型聚类的案例，Q 型聚类与 R 型聚类的过程相似，只不过在图 9-12 的对话框“分群”单选组中选择“个案”，即按样本的个案进行聚类，请读者根据具体问题选择不同的方法。

9.5 判 别 分 析

9.5.1 基本概念及统计原理

1. 基本概念

判别分析 (Discriminant Analysis) 是多元统计分析中用于判别样本所属类型的一种统计方法。它要解决的问题是在研究对象用某种方法已分成若干类的情况下, 确定新的观察数据属于已知类别中的哪一类。判别分析是应用很强的一种多元统计分析方法。如在经济学中, 根据国民收入、人均工农业产值、人均消费水平等多种指标来判定一个国家的经济发展程度所属类型。医生对病人病情的诊断, 需要根据观测到的病症 (如体温、血压、白血球数等) 判断病人患何种病等。

判别分析的假设: (1) 观测变量服从正态分布; (2) 观测变量之间没有显著的相关性; (3) 观测变量的平均值与方差不相关; (4) 观测变量应是连续变量, 因变量 (类别或组别) 是间断变量; (5) 两个观测变量的相关性在不同类中是一样的。

在判别分析的各阶段应把握以下原则: (1) 事前组别 (类) 的分类标准 (作为判别分析的因变量) 要尽可能地准确和可靠, 否则会影响判别函数的准确性, 从而影响判别分析的效果; (2) 所分析的自变量应是因变量的重要影响因素, 应该挑选既有重要特性又有区别能力的变量, 达到以最少的变量实现高辨别能力的目的; (3) 初始分析数据 (作为训练集的个案数) 不能太少。

2. 统计原理

判别分析按判别组数来分, 有两组判别分析和多组判别分析; 按区分不同总体所用的数学模型来分, 有线性判别和非线性判别。判别分析可以从不同角度提出问题, 因此有不同的判别准则, 如费希尔 (Fisher) 准则和贝叶斯 (Bayes) 准则。

判别分析用统计模型的语言来描述就是, 设有 m 个类 G_1, G_2, \dots, G_m , 希望建立一个准则, 对给定的任意一个样本 x , 依据这个准则就能判断它来自哪个类别。当然, 应当要求这种准则在某种意义下是最优的, 例如, 错判概率最小或错判损失最小等。

判别函数的一般形式是

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n \tag{9.2}$$

式中, y 为判别指标 (判别值); x_1, x_2, \dots, x_n 为反映研究对象特征的变量; a_1, a_2, \dots, a_n 为各变量的系数, 也称为判别系数。其中, 判别函数的个数为 \min (类别数-1, 预测变量数) 的值, 即类别数减 1 和观测变量数两个值之中的较小者。

3. 分析步骤

第 1 步 计算特征值。

计算需要用到的一些反映样本的特征值, 比如均值、协方差矩阵等。

第 2 步 建立判别函数。

判别函数的一般形式如式 (9.2), 建立判别函数就是要确定这些系数。

第 3 步 确定判别准则。

如费希尔准则和贝叶斯准则。

第 4 步 检验判别效果。

验证判别函数用来进行判别时的准确度。

第 5 步 分类。

根据所建立的判别函数对待判样本进行分类。SPSS 对于分成 m 个类的研究对象，建立 m 个线性判别函数。对于每个个体进行判别时，把测试的各变量值代入判别函数，得出判别得分，从而确定该个体属于哪一类（属于判别得分大的一类）；或者计算属于各类别的概率，从而判断该个体属于哪一类（属于概率最大的那一类）。

9.5.2 判别分析 SPSS 实例分析

【例 9-4】表 9.9 是健康人（ $c=1$ ）、硬化症患者（ $c=2$ ）和冠心病患者（ $c=3$ ）三种人群的心电图的 5 个指标（ $x_1 \sim x_5$ ）数据，其中有 19 个样本是确定的分类，另又测出 4 个人的相关指标，试根据确定分类的样本对未确定的样本进行分类。（参见数据文件：data9-4.sav。）

表 9.9 心电图测试数据

序号	x_1	x_2	x_3	x_4	x_5	c	序号	x_1	x_2	x_3	x_4	x_5	c
1	8.11	261.01	13.23	5.46	7.36	1	13	3.71	316.12	17.12	6.04	8.17	2
2	9.36	185.39	9.02	5.66	5.99	1	14	5.37	274.57	16.75	4.98	9.67	2
3	9.85	249.58	15.61	6.06	6.11	1	15	9.89	409.42	19.47	5.19	10.49	2
4	2.55	137.13	9.21	6.11	4.35	1	16	5.22	330.34	18.19	4.96	9.61	3
5	6.01	231.34	14.27	5.21	8.79	1	17	4.71	352.5	20.79	5.07	11	3
6	9.64	231.38	13.03	4.86	8.53	1	18	3.36	347.31	17.9	4.65	11.19	3
7	4.11	260.25	14.72	5.36	10.02	1	19	8.27	189.59	12.74	5.46	6.94	3
8	8.9	259.51	14.16	4.91	9.79	1	20	7.71	273.84	16.01	5.15	8.79	待定
9	8.06	231.03	14.41	5.72	6.15	1	21	7.51	303.59	19.14	5.7	8.53	待定
10	6.8	308.9	15.11	5.52	8.49	2	22	8.1	476.69	7.38	5.32	11.32	待定
11	8.68	258.69	14.02	4.79	7.16	2	23	4.71	331.47	21.26	4.3	13.72	待定
12	5.67	355.54	15.13	4.97	9.43	2							

第 1 步 分析。

由于部分样本已经有分类标记，还有几个待分类样本。这显然属于根据已知分类样本的信息对未分类样本进行分类的情况，用判别分析进行处理。

第 2 步 数据组织。

按表 9.9 所示，建立 7 个变量。分别是“序号”、“ x_1 ”、“ x_2 ”、“ x_3 ”、“ x_4 ”、“ x_5 ”和“ c ”，均为数值型变量。输入数据，对第 20 条~第 23 条的类别“ c ”变量，不填数据，作为缺失值处理并保存。

第 3 步 判别分析设置。

（1）按“分析→分类→判别式”顺序打开“判别分析”对话框，并按图 9-18 所示进行设置，对话框中的各项解释如下。

- ① “分组变量”框：选择类别变量，并单击下面的“定义范围(D)...”按钮，在“最小值”和“最大值”中分别输入分类变量的最小值和最大值。
- ② “自变量”框：选择参与判别分析的因素变量（自变量），即哪些因素决定了对分类的影响，下有两个单选按钮。
 - “一起输入自变量”选项：建立所选择的全部变量的判别式，这是系统默认的选项。
 - “使用逐步法”选项：采用逐步判别法进行判别分析。逐步判别法的基本思想与逐步回归一样，每一步选择一个判别能力最显著的变量进入判别函数，而且每次在选入变量之前对已进入判别函数的变量逐个进行检验。当每个变量因新变量的进入变得不显著时，

就将这个变量移出，直到判别函数中全部为有显著判别能力的变量。当发现自变量的判别能力有显著差异时，可考虑选择这个选项，将判别能力显著的变量“筛选”出来，建立“最优”的判别函数。这种方法有利于提高判别函数的判别能力。只有选择了此类方法后，“方法 (M) ...”按钮才被激活。

③ “选择变量”框：用于定义变量选择条件。选入变量以后，单击“值 (V) ...”按钮，弹出一个“设置值”子对话框，在对话框内输入一个数，表示全部记录中只有该变量取值等于这个数的记录才用于分析。

(2) “统计”对话框设置：单击“统计 (S) ...”按钮，弹出此子对话框，设置如图 9-19 所示，现对其中各选项解释如下。



图 9-18 “判别分析”对话框

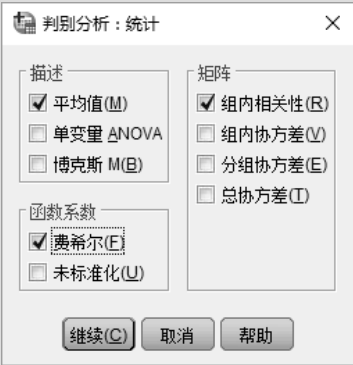


图 9-19 “判别分析：统计”对话框

- ① “描述”选项组：选择对原始数据的描述统计量的输出。
 - “平均值”选项：输出各类中各自变量的均值、标准差和各自变量总样本的均值、标准差。
 - “单变量 ANOVA”选项：对各类中同一自变量均值都相等的假设进行检验，输出单变量的分析结果。
 - “博克斯”选项：输出对各类协方差矩阵相等的假设进行 Box's M 检验的结果。
 - ② “函数系数”选项组：在默认情况下，系统给出的是采用 Bayes 方法建立的判别函数的标准化系数。
 - “费希尔”选项：Fisher 判别函数系数。可直接用于对新样本的分类，对每一类都给出一组系数，并且指出该类中具有最大判别分数的观测量。
 - “未标准化”选项：非标准化的判别函数系数。
 - ③ “矩阵”选项组：输出相关的矩阵。包括组内相关性矩阵、组内协方差矩阵、分组协方差矩阵和总协方差矩阵 4 项。
- (3) “分类”对话框的设置：单击“分类 (C) ...”按钮，弹出此子对话框，设置如图 9-20 所示，现对其中各选项解释如下。
- ① “先验概率”选项组：用于设定判别函数的先验概率。系统默认选中第一个选项“所有组相等”，即各类先验概率均相等，也就是各类平均分布；第二项“根据组大小计算”是指基于各类样本量占总样本的比例计算出先验概率，一般需选择该项。
 - ② “使用协方差矩阵”选项组：可选择“组内”，即为使用合并组内协方差矩阵进行分类，这是默认选项。“分组”表示为使用各组协方差矩阵进行分类。
 - ③ “显示”选项组：对需要输出的信息进行选择。

- “个案结果”选项：输出每个观测量的实际类、预测类、后验概率及判别分数。选中此项后，“将个案限制为前”被激活，可设置对前面 n 项观测量输出分类结果。
 - “摘要表”选项：输出分类小结表，对每一类输出判定正确和错判的观测量数。
 - “留一分类”选项：对于每一个观测量，输出依据除它之外的其他观测量导出的判别函数的分类结果。
 - ④ “图”选项组：对需要输出的图形进行选择。
 - “合并组”选项：生成包括各类的散点图，如果只有一个判别函数，则输出直方图。
 - “分组”选项：对每一类生成一张散点图，该图是根据前两个判别函数值作的。如果只有一个判别函数，则显示直方图。
 - “领域图”选项：根据判别函数值生成将观测变量分到各类去的边界图。图中每一类占一个区域，各类的均值用“*”标记出来。如果只有一个判别函数，则不显示此图。
- (4) “保存”对话框的设置：单击“保存(S)…”按钮，弹出此子对话框，设置如图 9-21 所示，现对其中各选项解释如下。



图 9-20 “判别分析：分类”对话框



图 9-21 “判别分析：保存”对话框

- ① “预测组成员”选项：建立新变量（默认变量名为 Dis_1），保存预测观测量所属类的值。
 - ② “判别得分”选项：建立新变量，保存判别分数。
 - ③ “组成员概率”选项：建立新变量，保存各观测量属于各类的概率值。
- 各项设置完成后提交系统运行。

第 4 步 主要结果及分析。

主要结果如表 9.10~表 9.16 及图 9-22、图 9-23 所示，具体分析如下。

(1) 表 9.10 是分类处理案例摘要表，表明共 23 条记录，已分好类的有 19 条（用这 19 个样本进行学习得到相应的分类概率），有 4 条需进行分类。

表 9.10 分析案例处理摘要

未加权个案数		个案数	百分比
有效		19	82.6
排除	缺失或超出范围组代码	4	17.4
	至少一个缺失判别变量	0	.0
	既包括缺失或超出范围组代码，也包括至少一个缺失判别变量	0	.0
	总计	4	17.4
总计		23	100.0

(2) 表 9.11 给出了这 5 个自变量之间的相关系数, 如变量 “ x_1 ” 与变量 “ x_2 ” 之间的相关系数为 0.059。

表 9.11 汇聚组内矩阵

		x1	x2	x3	x4	x5
相关性	x1	1.000	.059	-.008	-.203	-.090
	x2	.059	1.000	.835	-.328	.762
	x3	-.008	.835	1.000	-.187	.688
	x4	-.203	-.328	-.187	1.000	-.659
	x5	-.090	.762	.688	-.659	1.000

(3) 表 9.12 是特征值表, 由于本例中预测变量为 5 个, 类别数为 3, 因此判别函数的个数为 2 (即 $\min(3-1, 5)=2$)。判别函数的特征值越大, 表明该函数越具有区别力。第一个判别函数的特征值为 1.386, 第二个为 0.408。

表 9.12 特征值

函数	特征值	方差百分比	累计百分比	典型相关性
1	1.386 ^a	77.3	77.3	.762
2	.408 ^a	22.7	100.0	.538

a. 在分析中使用了前 2 个典则判别函数。

(4) 表 9.13 是对判别函数的显著性检验结果表, 其中 “1 直至 2” 表示两个判别函数的平均数在 3 个级别间的差异情况。 “2” 表示在排除第一个判别函数后, 第二个函数在 3 个级别间的差异情况。从最后的显著性概率来看, 这两个判别函数的效果并不十分显著。

表 9.13 威尔克 Lambda

函数检验	威尔克 Lambda	卡方	自由度	显著性
1 直至 2	.298	16.962	10	.075
2	.710	4.787	4	.310

(5) 表 9.14 为标准化的典则判别函数系数表, 根据此表可得判别函数:

$$F_1 = 0.626x_1 - 0.988x_2 - 0.664x_3 + 0.974x_4 + 1.434x_5$$

$$F_2 = 0.234x_1 + 1.808x_2 - 1.398x_3 + 0.416x_4 - 0.336x_5$$

根据这两个判别函数, 代入各变量的值可以计算出判别分数。根据各观测量的两个判别分数可以画出区域图或散点图, 具体如图 9-22 所示。

(6) 表 9.15 是分类处理摘要表。从表中可以看出, 有 23 条个案被成功分类。

表 9.14 标准化典则判别函数系数

	函数	
	1	2
x1	.626	.234
x2	-.988	1.808
x3	-.664	-1.398
x4	.974	.416
x5	1.434	-.336

表 9.15 分类处理摘要

已处理		23
排除	缺失或超出范围组代码	0
	至少一个缺失判别变量	0
已在输出中使用		23

(7) 表 9.16 为分类函数系数表, 根据该表可建立三个分类函数。

$$q_1 = 7.360x_1 - 0.222x_2 - 5.354x_3 + 104.590x_4 + 30.920x_5 - 369.692$$
$$q_2 = 6.891x_1 - 0.160x_2 - 5.209x_3 + 100.626x_4 + 29.073x_5 - 349.655$$
$$q_3 = 6.681x_1 - 0.211x_2 - 4.227x_3 + 98.616x_4 + 29.230x_5 - 340.370$$

将各变量值代入这三个判别函数模型进行计算, 对三者数进行比较, 将每个样本分到数值较大的类中。

(8) 图 9-22 是各类区域图及分类标记情况。这是以根据每个个案计算出的判别分数为坐标, 以典则判别函数 1 为横轴, 以典则判别函数 2 为纵轴, 所绘出的散点图。可以看出, 在图中分出了 1, 2, 3 三个区域。其中 1 表示“健康”, 2 表示“硬化病”, 3 表示“冠心病”, 在图中标出了各类的中心 (其中心用 “*” 表示)。

表 9.16 分类函数系数

	c		
	健康	硬化病	冠心病
x1	7.360	6.891	6.681
x2	-.222	-.160	-.211
x3	-5.354	-5.209	-4.227
x4	104.590	100.626	98.616
x5	30.920	29.073	29.230
(常量)	-369.692	-349.655	-340.370

费希尔线性判别函数

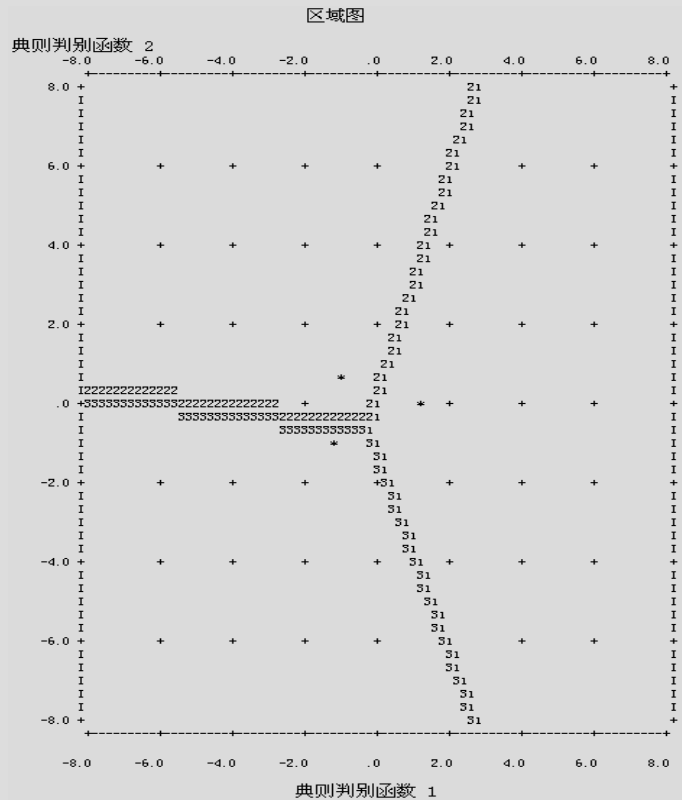


图 9-22 各类区域图及其标记说明

(9) 表 9.17 是分类结果表。对角线显示的为准确预测的个数, 其余为错误预测的个数。从该表可以看出, 已经分类的 19 个个案中正确分类 17 个, 错误分类 2 个。正确率比较高。以这 19 个个案为先验数据, 将待分类的 4 个个案分别分入 1, 2, 3 类的分别有 1, 1, 2 个。

表 9.17 分类结果^a

		c	预测组成员信息			总计
			健康	硬化病	冠心病	
原始	计数	健康	9	0	0	9
		硬化病	0	5	1	6
		冠心病	1	0	3	4
		未分组个案	1	1	2	4
	%	健康	100.0	.0	.0	100.0
		硬化病	.0	83.3	16.7	100.0
		冠心病	25.0	.0	75.0	100.0
		未分组个案	25.0	25.0	50.0	100.0

a. 正确地对 89.5% 个原始已分组个案进行了分类。

(10) 图 9-23 是工作数据文件中的记录情况。由于选择了保存“预测组成员”，即要求保存各个个案的分类情况，可以看出，数据文件中增加了一列“Dis_1”，记录对应的分类情况。

序号	x1	x2	x3	x4	x5	c	Dis_1
1	8.11	261.01	13.23	5.46	7.36	1	1
2	9.36	185.39	9.02	5.66	5.99	1	1
3	9.85	249.58	15.61	6.06	6.11	1	1
4	2.55	137.13	9.21	6.11	4.35	1	1
5	6.01	231.34	14.27	5.21	8.79	1	1
6	9.64	231.38	13.03	4.86	8.53	1	1
7	4.11	260.25	14.72	5.36	10.02	1	1
8	8.90	259.51	14.16	4.91	9.79	1	1
9	8.06	231.03	14.41	5.72	6.15	1	1
10	6.80	308.90	15.11	5.52	8.49	2	2
11	8.68	258.69	14.02	4.79	7.16	2	2
12	5.67	355.54	15.13	4.97	9.43	2	2
13	3.71	316.12	17.12	6.04	8.17	2	2
14	5.37	274.67	16.75	4.98	9.67	2	3
15	9.89	409.42	19.47	5.19	10.49	2	2
16	5.22	330.34	18.19	4.96	9.61	3	3
17	4.71	352.50	20.79	5.07	11.00	3	3
18	3.36	347.31	17.90	4.65	11.19	3	3
19	8.27	189.59	12.74	5.46	6.94	3	1
20	7.71	273.84	16.01	5.15	8.79	.	1
21	7.51	303.59	19.14	5.70	8.53	.	3
22	8.10	476.69	7.38	5.32	11.32	.	2
23	4.71	331.47	21.26	4.30	13.72	.	3

图 9-23 数据文件中分类结果

☆说明☆

◆ 为了简化起见，例 9-4 将所有自变量全部进入模型进行判别分析（见图 9-18），读者也可选择“使用步进法”，其原理与回归分析中的“步进”类似。当选择此方法后，“方法（M...）”按钮被激活，与“逐步”回归方法类似，可选择相应的方法，具体操作请读者自己练习。

9.6 典型案例

9.6.1 美国 22 家企业类型划分

美国一家咨询公司为了研究不同类型企业的特征，分别调查了亚得桑那公共服务公司、爱迪生公司、联合装饰公司、维多利亚电力公司等 22 家企业，并收集了 5 个指标，即 x_1 ：固定支出综合率（%）， x_2 ：资产收益率（%）， x_3 ：每千瓦成本（美元）， x_4 ：每年使用的能源（万千瓦时），

x_5 : 是否使用核能源（其中，0 表示没有使用核能，1 表示使用了核能）。其具体数据如表 9.18 所示。试根据这些统计指标对这 22 家企业进行聚类，并观测不同类型的企业所具有的特征。（数据来源：杨维忠等，《SPSS 统计分析 with 行业应用案例详解》，清华大学出版社；参见数据文件：data9-5.sav。）

表 9.18 美国 22 家企业的统计数据

编号	x_1	x_2	x_3	x_4	x_5	编号	x_1	x_2	x_3	x_4	x_5
1	1.06	9.2	351	9077	0	12	1.13	6.3	457	6154	0
2	0.89	13.6	202	5088	1	13	0.63	12.7	199	1175	1
3	1.43	8.9	521	9212	0	14	1.09	6.1	296	9673	0
4	0.78	11.2	168	6423	1	15	0.96	17.6	164	2468	1
5	0.66	16.3	192	3300	1	16	1.16	9.9	252	15991	0
6	0.75	13.5	111	1127	1	17	0.76	16.4	136	4714	1
7	1.22	3.6	1705	7642	0	18	1.05	2.8	351	10140	0
8	1.1	9.2	245	13082	0	19	1.16	4.9	401	13507	0
9	1.34	13	456	8406	0	20	0.48	11.8	148	2279	1
10	0.58	12.4	197	3455	1	21	1.04	8.4	442	6650	0
11	1.25	7.5	376	17441	0	22	0.36	16.3	184	1093	1

案例分析：这是一个典型的聚类问题，是对个案（企业）的聚类，共有 5 个变量，其中有 4 个连续型变量（ $x_1 \sim x_4$ ），1 个类型变量（ x_5 ）。根据各种聚类方法适用的特点，可选择使用二阶聚类进行分析，在聚类分析的基础上再分析不同类型企业在各指标上具有的特征值。

9.6.2 销售地区的选择

湖南省某白酒厂开发了一种新的“白酒”，想在本省上市，考虑到公司的现状：生产能力小，营销实力不强，在全省范围内没有系统的营销网络。公司收集了 2015 年度湖南省各地区的经济发展和消费水平指标，并选取了与白酒消费相关的 6 个代表性指标，即 x_1 : 总人口数（万人）， x_2 : 人均地区生产总值（元）， x_3 : 在岗职工年均工资（元）， x_4 : 农村居民人均可支配收入（元）， x_5 : 城镇居民人均可支配收入（元）， x_6 : 在职职工人数（万人）。具体数据如表 9.19 所示，试根据该厂的特点选择营销区域。（数据来源：《湖南统计年鉴》，2016 年；参见数据文件：data9-6.sav。）

表 9.19 湖南 14 个地区的经济发展及消费情况指标

地区	x_1	x_2	x_3	x_4	x_5	x_6
长沙	743.18	114509.7	67266	23601	39961	123.27
株洲	400.05	58370.45	57584	15637	33977	42.81
湘潭	282.37	60314.48	51742	15347	29237	26.69
衡阳	733.75	35455.95	44983	14407	26515	50.65
邵阳	726.17	19100.21	47249	8716	21070	34.70
岳阳	562.92	51273.36	45592	12091	25202	43.00
常德	584.39	46356.37	48655	11744	24513	39.06
张家界	152.40	29376.64	48425	7094	19473	7.85
益阳	441.02	30710.85	48724	12344	22571	23.03
郴州	473.02	42536.68	49086	11778	25534	32.32
永州	542.97	26118.94	44553	10765	21938	29.90
怀化	490.16	25976.21	49006	7203	20693	24.43
娄底	387.18	33360.71	43986	8655	21838	27.66
湘西	263.45	18881.76	48226	6648	19267	13.06

案例分析：由于该公司是一家小型企业，经济实力不够强大，如果要在全省范围内全面上市，需投入太多的成本，而且会面临很大的风险。最好的方式是，公司根据企业自身的实际情况和产品的特点，先选择与居民消费能力相当的几个地区进行试销，如果有效，再扩大到其他地区。于是就需根据各地区的经济状况和居民消费水平，将 14 个地区进行聚类，此问题就是一个聚类问题，可采用二阶聚类、K-均值聚类和系统聚类等方法进行分析。根据聚类的结果，公司可选择进入市场的相应策略。

9.6.3 地区降水量区域类型判别

我国华北地区和长江中下游地区的降水变化有不同特点，表 9.20 给出了华北地区和长江中下游地区一些观测站记录到的六月降水天数（rainday6）、八月降水天数（rainday8）、八月与六月降水量之比（ratio）的数据资料，同时给出了两地区中间地带一些观测站记录的相应观测数据。试判断这些中间地带的地区各与哪个区域的降水更相似。（数据来源：郝黎仁等，《SPSS 实用统计分析》；参见数据文件：data9-7.sav。）

表 9.20 各地区的降水情况

	编号	地区	rainday6	rainday8	ratio	区划类型
华北地区	1	北京	9.7	14.3	3.46	1
	2	天津	8.9	12.1	2.45	1
	3	保定	9	12.5	3.26	1
	4	石家庄	8.5	13	3.39	1
	5	太原	10.6	13.3	2.13	1
	6	大同	11.6	12.7	2.05	1
	7	张家口	11.4	12.7	1.82	1
	8	榆林	7.8	12.5	1.82	1
	9	兴县	10.1	13.3	3.01	1
	10	五台山	16.4	18.1	1.8	1
长江中下游地区	11	上海	13.1	10	0.74	2
	12	南京	10.9	11.5	0.87	2
	13	合肥	10.3	10.1	1.18	2
	14	汉口	11.7	8.5	0.61	2
	15	九江	13.6	9.4	0.61	2
	16	安庆	12.3	9.5	0.44	2
	17	芜湖	10.5	10.9	0.76	2
	18	墨阳	11.3	12.2	0.75	2
	19	黄石	14	10.4	0.64	2
	20	东山	12.5	11.7	1.01	2
待判	21	青岛	13.7	11.6	1.68	
	22	崇州	10.5	13.7	1.75	
	23	临沂	10	12	1.65	
	24	徐州	8.3	11.1	1.48	
	25	阜阳	8.6	10.9	1.07	

案例分析：要研究这 5 个未划分区域的地区与上述两区域之间的相似性，即可认为将这 5 个地区分类（判别）到这两个区域，这是一个典型的判别分析。由于编号为 1~20 的地区已经有明

确的分类，分别归到华北地区和长江中下游各地区，可先用这 20 个地区的数据建立判别函数，然后根据建立的判别函数将这 5 个地区划分到两个区域。

9.7 思考与练习

1. 聚类分析的意义和作用是什么？
2. 如何解读系统聚类后 SPSS 输出的聚类树状图和冰柱图？
3. 为了对游泳运动员进行分类，预计分为蝶泳、仰泳、蛙泳和自由泳 4 类，为简化问题，仅以 10 名运动员的三项测试数据为例。其中变量为 x_1 （肩宽/髋宽 $\times 100$ ）， x_2 （胸厚/胸围 $\times 100$ ）， x_3 （腿长/身长 $\times 100$ ），数据如表 9.21 所示，试进行聚类分析。（参见数据文件：data9-8.sav。）

表 9.21 游泳运动员的体测数据

no	1	2	3	4	5	6	7	8	9	10
x_1	125	121	120	124	122	120	121	122	122	121
x_2	20	18	17	20	18	19	17	19	17	19
x_3	44	43	42	45	43	44	41	43	42	45

4. 2015 年各地区的客运人数（单位：万人）如表 9.22 所示，其中包含铁路、公路和水运情况。试分别用二阶聚类和系统聚类对各地区的运输能力进行聚类分析。（数据来源：《中国统计年鉴》，2016 年；参见数据文件：data9-9.sav。）

表 9.22 2015 年全国各地区客运人数（单位：万人）

地区	铁路	公路	水运	地区	铁路	公路	水运
北京	12918	49931	0	湖北	13508	87953	574
天津	4054	14219	72	湖南	10511	119266	1534
河北	9706	43563	5	广东	23149	98050	2727
山西	7418	22085	109	广西	7046	41522	533
内蒙古	5108	11017	0	海南	1651	10363	1714
辽宁	12919	60269	504	重庆	3994	57556	732
吉林	7158	29013	188	四川	9207	124014	2748
黑龙江	9865	32632	372	贵州	4901	80621	2019
上海	9692	3766	386	云南	3949	43688	1157
江苏	16116	119800	2392	西藏	221	871	0
浙江	15224	92304	3841	陕西	7866	61436	378
安徽	8553	78072	185	甘肃	3123	37240	90
福建	9256	40394	1996	青海	936	4596	70
江西	8458	53687	273	宁夏	661	8444	195
山东	11397	46960	1999	新疆	2719	33229	0
河南	12200	112535	280				

5. 为了研究某地区育龄妇女的生育情况，根据生育峰值年龄、一胎生育率、二胎生育率、多胎生育率和总和生育率 5 项指标，收集到 12 个样品的分类情况，另收集到 3 个待判样品情况，数据如表 9.23 所示。（数据来源：罗积玉，《经济统计分析方法及预测》，清华大学出版社；参见数据文件：data9-10.sav。）

根据表中数据回答如下问题：

- （1）试用自变量全进入判别法和逐步判别法进行判别分析，决定三个待判样本应归属哪一类，并比较二者的差异。

(2) 使用逐步判别法进行分析时, 在“方法”对话框中改变“方法”栏中的方法和“标准”栏中的“进入”值和“删除”值, 观察对判别结果的影响。

表 9.23 生育情况数据

序号	峰值年龄	一胎生育率	二胎生育率	三胎生育率	组别
1	27	96.77	2.8	0.43	1
2	24	55.33	25.36	19.31	1
3	27	97.45	2.1	0.45	1
4	24	51.45	31.25	17.3	1
5	25	52.15	32.85	16	1
6	25	52.08	32.84	15.08	1
7	25	35.75	22.83	41.41	2
8	26	27.1	25.13	47.77	2
9	25	39.4	34.21	26.39	2
10	26	21.98	16.23	61.79	2
11	25	38.49	34.44	27.06	2
12	25	38.96	24.48	36.56	2
13	26	87.45	12.5	0.05	待测
14	25	33.78	22.82	43.4	待测
15	24	52.4	33.25	14.35	待测

6. 我国 2012 年各地城镇居民平均全年家庭收入来源统计如表 9.24 所示, 试对全国各地的城镇居民家庭收入来源结构进行分类。(数据来源:《中国统计年鉴》, 2013 年; 参见数据文件: data9-11.sav。)

表 9.24 2012 年各地区城镇居民全年家庭收入统计表 (单位: 元)

地区	工薪收入	经营净收入	财产性收入	转移性收入
北京	27691.8	1430.2	717.6	10993.5
天津	21523.8	1200.1	515.5	9704.6
河北	13154.5	2257.5	338.5	6149.0
山西	14973.6	1041.4	301.8	5783.4
内蒙古	16872.6	2698.7	564.0	4655.5
辽宁	14846.1	2710.3	493.0	7866.4
吉林	13535.3	2168.8	324.0	5631.5
黑龙江	11700.5	1729.3	186.1	5752.0
上海	31109.3	2267.2	575.8	10802.2
江苏	26102.1	3421.9	690.0	8305.2
浙江	22385.1	4694.4	1465.3	9450.0
安徽	14812.5	2155.3	549.6	6007.1
福建	19976.0	3337.0	1795.2	5769.7
江西	13348.1	1946.8	527.6	5327.7
山东	19856.1	2621.4	704.9	4823.2
河南	13666.5	2545.1	333.8	5351.8
湖北	14191.0	2158.3	476.2	6078.3
湖南	13237.1	3008.3	867.8	5691.4
广东	23632.2	3603.9	1468.7	5339.6

续表				
地区	工薪收入	经营净收入	财产性收入	转移性收入
广西	14693.5	2131.8	883.7	5500.4
海南	14672.3	2397.4	717.6	5022.5
重庆	15415.4	2183.5	538.4	6673.6
四川	14249.3	2017.8	633.8	5427.3
贵州	12309.2	1982.5	355.7	5395.6
云南	14408.3	2425.0	1000.0	5167.1
西藏	17672.1	570.9	417.9	1563.3
陕西	15547.3	882.0	269.6	5907.1
甘肃	12514.9	1125.7	259.6	5098.2
青海	12614.4	1191.4	93.0	5847.8
宁夏	13965.6	2522.8	160.9	5252.9
新疆	14432.1	1633.2	145.5	3983.7

7. 为明确诊断出小儿肺炎三种类型，某研究单位测得 30 名结核性、12 名化脓性和 18 名细菌性肺炎患儿共 60 名的 7 项生理、生化指标，试建立判别函数。（数据来源：武松 等，《SPSS 统计分析大全》，清华大学出版社；参见数据文件：data9-12.sav。）

第 10 章 主成分分析和因子分析

在科学研究中，往往需要对反映事物的多个变量进行大量观测，收集大量数据以便进行分析，寻找规律。例如，对高等学校科研状况的评价研究，可能会收集诸如投入科研活动的人数、立项课题数、项目经费、经费支出、结项课题数、发表论文数、发表专著数、获得奖励数等多项指标。多变量大样本无疑会对科学研究提供丰富的信息，但也在一定程度上增加了数据采集的工作量，更重要的是，在大多数情况下，许多变量之间可能存在相关性而增加了问题分析的复杂性，在实际建模时，这些变量未必能真正发挥预期的作用，有些变量的存在可能反而给问题的分析带来许多问题。如果分别分析每个指标，分析又可能是孤立的，而不是综合的。盲目减少指标会损失很多信息，容易产生错误的结论。因此需要找到一个合理的方法，减少分析指标的同时，尽量减少原指标包含信息的损失，对所收集的资料做全面分析。由于各变量间存在一定的相关性，因此有可能用较少的综合指标分别综合存在于各变量中的各类信息。在 SPSS 中进行因子分析和主成分分析，可以利用“分析→降维→因子”过程来实现。

10.1 主成分分析和因子分析简介

主成分分析的目的是用较少的变量去解释原始数据中的大部分变异，这些变量也就是利用主成分分析法整理而成的整体性指标。而因子分析也是希望能够降低变量的数目，但不同的是，只想在一群具有相关性且难以解释的数据中，找出几个在概念上有意义的，并且彼此之间近于独立的，可以影响原始数据的共同因素。主成分分析和因子分析实际上都是“降维”方法，首先对它们所涉及共同概念介绍如下。

10.1.1 基本概念和主要用途

1. 基本概念

主成分分析（Principal Component Analysis）就是考虑各指标之间的相互关系，利用降维的方法将多个指标转换为少数几个互不相关的指标，从而使进一步研究变得简单的一种统计方法。主成分分析是由 Hotelling 于 1933 年首先提出的，是利用“降维”的思想，在损失很少信息的前提下把多个指标转化为几个综合指标，称为主成分。每个主成分均是原始变量的线性组合，且各个主成分之间互不相关，这就使得主成分比原始变量具有某些更优越的性能。主成分分析结果不能看成研究的结果，而应该在主成分分析的基础上继续采用其他多元统计方法来解决实际问题。

因子分析是一种通过显在变量测评潜在变量，通过具体指标测评抽象因子的分析方法，最早是由心理学家 Chales Spearman 在 1904 年提出的，它的基本思想是将实测的多个指标，用少数几个潜在指标（因子）的线性组合表示。因子分析主要应用到两个方面：一是寻求基本结构，简化观测系统；二是对变量或样本进行分类。

因子分析的基本思想是根据相关性的大小把变量分组，使得同组内的变量相关性较高，而不同组的变量相关性较低。每组变量代表一个基本结构，这个基本结构称为一个公共因子。对于所研究的问题就可试图用最少数不可测的公共因子的线性函数与特殊因子之和来描述原来观测的

每一个分量。因子分析还可以用于对变量或样本的分类处理，可根据因子得分值，在因子轴所绘制的空间中把变量或样本点画出来，形象直观地达到分类的目的。通常将研究变量之间相互关系的因子分析称为 R 型因子分析，而将研究样本之间相互关系的因子分析称为 Q 型因子分析。本节将重点介绍 R 型因子分析。

2. 主要用途

(1) 解决共线性问题：利用主成分分析提取出主要信息，然后使用提取出的主成分代替原变量进行分析，就可以避开原变量的共线性问题。

(2) 评估问卷的结构效度：运用因子分析得出问卷中哪些问题用于研究那些潜在的特征（因子），从而得出对该问卷结构效度的评价。这是社会学和流行病学调查中常用的方法。

(3) 寻找变量之间的潜在结构：许多变量是无法直接观测到的，它们往往需要用一系列可直接观测的相关变量来间接反映。运用因子分析，就可以将这些变量潜在的结构推导出来并加以利用。

(4) 内在结构证实：在某些情况下，研究者根据某些理论或其他知识对可能的内在结构进行了假设，此时可利用因子分析来验证该假设是否成立，这种因子分析又称为证实性因子分析，在心理学研究中较为常见。

3. 常用术语

主成分分析和因子分析中常用的主要概念如下。

(1) 因子载荷

因子载荷 a_{ij} 就是第 i 个变量与第 j 个公共因子之间的相关系数，它的统计意义就是第 i 个变量在第 j 个公共因子上的负荷，反映了第 i 个变量在第 j 个公共因子上的相对重要性。

(2) 变量共同度

变量共同度，也称公共方差，反映全部公共因子变量对原有变量 x_i 的总方差的解释说明比例。原有变量 x_i 的共同度为因子载荷矩阵 A 中第 i 行元素的平方和，即

$$h_i^2 = \sum_{j=1}^m a_{ij}^2 \quad (10.1)$$

h_i^2 越接近于 1（原有变量 x_i 在标准化前提下，总方差为 1），说明公共因子解释原有变量的信息越多。通过该值，可以掌握该变量的信息有多少丢失了。如果大部分变量的共同度都高于 0.8，则说明提取出的公共因子已经基本反映了各原始变量 80% 以上的信息，仅有较少的信息丢失，因子分析的效果较好。

(3) 公共因子 F_j 的方差贡献

公共因子 F_j 的方差贡献定义为因子载荷矩阵 A 中第 j 列各元素的平方和，即

$$S_j = \sum_{i=1}^p a_{ij}^2 \quad (10.2)$$

公共因子 F_j 的方差贡献反映了该因子对所有原始变量总方差的解释能力，其值越大，说明因子重要程度越高。

10.1.2 主成分和公因子数量的确定

主成分分析（或因子分析）希望用尽可能少的主要成分（公共因子）包含原来尽可能多的信息，那么如何确定需要保留的主成分数量（或公共因子数量）呢？可以遵循以下几个原则。

- 主成分的累积贡献率：一般来说，提取主成分的累积贡献率达到 80%~85%以上就比较满意了，可以由此确定需要提取多少个主成分。
- 特征值：特征值在某种程度上可以看成表示主成分影响力度大小的指标，如果特征值小于 1，说明该主成分的解释力度还不如直接引入原变量的平均解释力度大。因此一般可以用特征值大于 1 作为纳入标准。
- 综合判断：大量的实际情况表明，如果根据累积贡献率来确定，主成分数往往较多，而用特征值来确定，又往往较少，很多时候应当将两者结合起来，以综合确定合适的数量。

10.1.3 主成分分析与因子分析的区别与联系

(1) 两者都是在多个原始变量中通过它们之间的内部相关性来获得新的变量（主成分变量或因子变量），达到既能减少分析指标个数，又能概括原始指标主要信息的目的。它们各有特点：主成分分析是将 m 个原始变量提取 k ($k \leq m$) 个互不相关的主成分；因子分析是提取 k ($k \leq m$) 个支配原始变量的公共因子和 1 个特殊因子，各公因子之间可以相关或互不相关。

(2) 提取公因子主要有主成分分析法和公因子法，若采用主成分法，则主成分分析和因子分析基本等价，该法主要从解释变量的总方差角度，尽量使变量的方差被主成分解释，即主成分分析方法倾向得到更大的共性方差，而公因子法主要从解释变量的相关性角度，尽量使变量的相关程度能被公因子解释，当因子分析的目的为确定结构时会用到该法。

(3) 因子分析提取的公因子比主成分分析提取的主成分更具有解释性。主成分分析不考虑观测变量的度量误差，直接用观测变量的某种线性组合来表示一个综合变量，而因子分析的潜在变量则校正了观测变量的度量误差，且它还可进行因子旋转，使潜在因子的实际意义更明确，分析结论更真实。

(4) 两者分析的实质和重点不同。主成分分析的模型为 $Y=BX$ ，即主成分 Y 为原始变量 X 的线性组合。因子分析的数学模型为 $X=BF+$ ，即原始变量 X 为公因子 F 与特殊因子的线性组合。可知，主成分分析主要是综合原始变量的信息，而因子分析重在解释原始变量之间的关系。主成分分析实质上是线性变换，无假设检验，而因子分析是统计模型，某些因子模型是可以得到假设检验的。

(5) 两者的 SPSS 操作都是通过“分析→降维→因子”过程实现的，主成分分析不需要因子旋转，而因子分析需要经过旋转。

10.2 主成分分析

10.2.1 基本概念及统计原理

1. 统计原理

定义 10.1 设随机向量 $\mathbf{x}'=(x_1,x_2,\cdots,x_p)$ 的相关系数矩阵为 \mathbf{R} （也可为协方差矩阵 Σ ）， $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ 为 \mathbf{R} 的特征值， $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p$ 为对应的标准正交特征向量。则第 i 个主成分为

$$y_i = \mathbf{e}_i' \mathbf{x} = e_{1i}x_1 + e_{2i}x_2 + \cdots + e_{pi}x_p, i = 1, 2, \cdots, p \tag{10.3}$$

此时有

$$\text{Var}(y_i) = \mathbf{e}_i' \mathbf{R} \mathbf{e}_i = \lambda_i, i = 1, 2, \cdots, p \tag{10.4}$$

$$\text{Cov}(y_i, y_k) = \mathbf{e}_i' \mathbf{R} \mathbf{e}_k = 0, i \neq k \quad (10.5)$$

若一些 λ_i 有重根, 则系数向量 \mathbf{e}_i 和 y_i 不唯一。

定义 10.1 中的标准化正交特征向量 $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$ 总是存在的。事实上, 若特征值 $\lambda_1, \lambda_2, \dots, \lambda_p$ 都不相等, 那么诸 \mathbf{e}_i 自然是正交的。若一些特征值有重根, 也能够选择对应于这些特征值的特征向量, 使得它们是正交的。定义 10.1 表明, x_1, x_2, \dots, x_p 的主成分是以 \mathbf{R} 的特征向量为系数的线性组合, 它们互不相关, 其方差为 \mathbf{R} 的特征值。

设第 k 个主成分的方差占总方差的比例为 p_k , 则有

$$p_k = \frac{\lambda_k}{\sum_{i=1}^p \lambda_i} \quad (10.6)$$

当变量个数 p 较大时, 如果前若干个主成分的方差之和占了总方差的很大一部分 (如 85% 以上), 用这些主成分代替原 p 个变量, 不会损失太多信息。系数向量 $\mathbf{e}_i' = (e_{i1}, e_{i2}, \dots, e_{ip})$ 的分量也有一定的意义。 e_{ki} 刻画了第 k 个变量对第 i 个主成分的重要性。主成分的计算公式为

$$\begin{cases} y_1 = e_{11}x_1 + e_{12}x_2 + \dots + e_{1m}x_m \\ y_2 = e_{21}x_1 + e_{22}x_2 + \dots + e_{2m}x_m \\ \vdots \\ y_p = e_{p1}x_1 + e_{p2}x_2 + \dots + e_{pm}x_m \end{cases} \quad (10.7)$$

2. 分析步骤

假定输入一个决策表 $T=(U, C \cup D, V, f)$, 其中 U 为论域, $X=\{x_1, x_2, \dots, x_m\}$, C 和 D 分别为条件属性集和决策属性集。需输出条件属性的主成分 $P=\{y_1, y_2, \dots, y_p\}$ 。则其步骤如下。

第 1 步 原始数据的标准化处理。

按 $x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\text{var}(x_j)}}$ 进行标准化处理, 使每个属性均值为 0, 方差为 1。

第 2 步 计算相关系数矩阵。

计算第 1 步中得到的数据集 X 的相关系数矩阵 \mathbf{R} 。

第 3 步 计算特征值及单位特征向量。

计算 \mathbf{R} 的特征值 λ_i 及其对应的单位特征向量 \mathbf{e}_i , $i=1, 2, \dots, m$, 并将特征值按由大到小的顺序排列, 即 $\lambda_1 > \lambda_2 > \dots > \lambda_m$ 。

第 4 步 计算主成分的方差贡献率和累积方差贡献率。

第 k 个主成分方差为 $a_k = \lambda_k / \left(\sum_{i=1}^m \lambda_i \right)$, 主成分 y_1, y_2, \dots, y_p 的累积方差贡献率为 $\left(\sum_{i=1}^p \lambda_i \right) / \left(\sum_{j=1}^m \lambda_j \right)$ 。

其中 a_1 的值最大, 说明 y_1 综合 x_1, x_2, \dots, x_m 信息的能力最强, 主成分 p 值的选取一般为使得累积方差贡献率 $\geq 80\%$ (或特征值大于 1) 的前 p 个特征值。

第 5 步 计算主成分。

利用前 p 个特征值对应的单位特征向量 $\mathbf{e}_1=(e_{11}, e_{12}, \dots, e_{1m})'$, $\mathbf{e}_2=(e_{21}, e_{22}, \dots, e_{2m})'$, \dots , $\mathbf{e}_m=(e_{m1}, e_{m2}, \dots, e_{pm})'$, 按式 (10.7) 计算原始数据的主成分 y_1, y_2, \dots, y_p 。

10.2.2 主成分分析 SPSS 实例分析

【例 10-1】 为了从总体上反映 20 世纪末世界经济全球化的状况，现选择了 1999 年全球具有代表性的 16 个国家的数据，这些国家参与经济全球化的程度指标值如表 10.1 所示（其各指标的具体含义如表 10.2 所示）。试分析一个国家参与经济全球化的程度主要受哪些因素的影响。（数据来源：刘玉玫，《经济全球化程度的量化研究》，统计研究；参见数据文件：data10-1.sav。）

表 10.1 部分国家参与经济全球化的指标数据 （单位：%）

编号	国家	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	x ₇	x ₈	x ₉	x ₁₀	x ₁₁	x ₁₂	x ₁₃	x ₁₄	x ₁₅
1	中国	3.205	54.5	28.53	0.878	1.409	0.894	11.6	2.305	0.547	2.932	4.818	9.003	2.7	3.914	1.472
2	印度	1.449	31.1	0.279	0.339	0.272	0.1	2.7	0.128	0.193	0.825	2.318	5.127	0.6	4	0.218
3	日本	14.079	52.3	0.653	10.254	11.769	1.097	0	1.967	1.3	6.178	14.746	27.297	30.9	57.734	15.125
4	韩国	1.318	136.3	1.011	1.6	0.42	1.838	1.3	0.77	0.78	2.267	23.32	42.875	9.1	12.129	0.452
5	新加坡	0.275	739.5	3.572	27.841	0.884	13.314	28.6	0.622	0.143	1.885	169.772	319.907	54.2	917.328	0.718
6	美国	29.641	46.1	3.682	6.429	20.563	4.808	5.4	24.253	29.941	15.638	10.784	24.555	13.6	24.495	21.274
7	加拿大	2.056	101.5	0.898	8.276	2.313	5.369	10.5	2.444	5.145	3.854	34.691	67.047	15.1	21.83	1.362
8	巴西	2.434	27.1	1.584	2.327	0.962	2.905	6.8	1.953	2.3	0.857	4.716	10.101	6.7	5.498	1.104
9	墨西哥	1.567	151.4	1.657	2.837	0.797	1.471	10.9	0.67	0.212	2.186	18.485	37.986	4.5	4.887	0.468
10	英国	4.67	118.4	0.497	26.151	12.456	22.137	11.2	16.552	19.642	5.542	28.434	58.7	66.1	278.968	11.289
11	法国	4.639	120.6	1.84	9.242	4.492	10.848	8.5	8.282	5.841	5.21	28.46	54.052	29.2	56.453	8.889
12	德国	6.84	132.9	2.252	9.558	6.646	7.747	2.2	8.589	8.971	8.843	32.121	63.174	36	51.514	12.18
13	意大利	3.792	104.5	0.321	8.153	3.724	1.059	2.5	0.77	1.913	4.032	22.869	43.924	27	17.776	5.678
14	俄罗斯	1.3	58.6	1.533	1.499	0.552	0.499	2.5	0.31	0.298	0.987	7.77	12.581	1.1	2.001	0.469
15	澳大利亚	1.309	94.5	0.502	5.773	0.941	1.987	18.9	0.527	1.371	1.131	15.745	33.795	13.2	24.117	0.797
16	新西兰	0.177	110.5	0.218	7.374	0.179	3.04	31.5	0.126	0.338	0.248	23.221	47.387	19.8	41.274	0.215

表 10.2 对指标意义的解释

指标	指标意义
x ₁	GDP 占全球 GDP 的比重
x ₂	货物贸易占货物 GDP 的比重
x ₃	外国分支机构占世界全部分支机构的比重
x ₄	本国发生的全部收益占 GDP 的比重
x ₅	本国发生的全部收益占世界发生的全部收益的比重
x ₆	对外直接投资和接受外国直接投资总额占 GDP 的比重
x ₇	外国直接投资占国内投资总额的比重
x ₈	本国直接投资额占全球直接投资额的比重
x ₉	跨国并购额占全球跨国并购额的比重
x ₁₀	国际经济外向度
x ₁₁	对外贸易依存度
x ₁₂	货物和服务进出口总额占 GDP 的比重
x ₁₃	国际金融总资本流量占 GDP 的比重
x ₁₄	对外金融资产负债占 GDP 的比重
x ₁₅	国际金融总资本流量占全球国际金融总资本流量的比重

第 1 步 分析。

从数据来看，一共有 15 个因素，但有些因素是存在相关性的，同时各因素对全球化影响的程度也是不一样的，故可采用主成分分析。

第 2 步 数据组织。

按表 10.2 中的“指标”一列定义变量，输入表 10.1 所示的数据并保存。

第 3 步 主成分分析的设置。

(1) 按“分析→降维→因子”顺序打开“因子”对话框，将 $x_1 \sim x_{15}$ 这 15 个变量移入“变量”对话框中，如图 10-1 所示。

现对对话框中的各项解释如下。

① “变量”框：选择用于进行因子分析或主成分分析的变量。

② “选择变量”框：用于定义变量选择条件。选入变量以后，单击“值(L)…”按钮，弹出一个“设置值”子对话框，在对话框内输入一个数，表示全部记录中只有该变量取值等于这个数的记录才用于分析。

(2) “描述”子对话框的设置：单击“描述(D)…”按钮，弹出此子对话框，设置如图 10-2 所示，现对其中各选项解释如下。



图 10-1 “因子分析”对话框

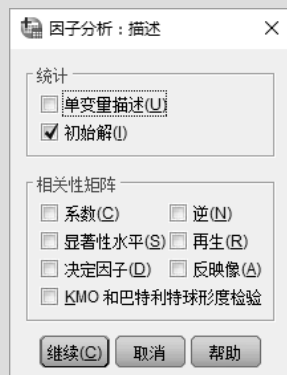


图 10-2 “因子分析：描述”对话框

① “统计”选项组。

- “单变量描述”选项：用于输出参与分析的原始变量的均值、标准差等描述统计量。
- “原始解”选项：给出因子提取前，分析变量的公因子方差。对主成分分析来说，这些值是分析变量的相关矩阵或协方差矩阵的对角元素。对因子分析来说，是每个变量用其他变量作预测因子的载荷平方和。

② “相关性矩阵”选项组。

- “系数”选项：给出原始分析变量间的相关系数矩阵。
- “显著性水平”选项：给出每个相关系数相对于 0 的单尾假设检验的显著性水平。
- “决定因子”选项：给出相关系数矩阵的行列式。
- “逆”选项：给出相关系数矩阵的逆矩阵。
- “再生”选项：再生相关矩阵。此项给出因子分析后的相关矩阵，还给出残差，即原始相关与再生相关之间的差值。
- “反映像”选项：给出反映像相关矩阵，包括偏相关系数的负数；反映像协方差矩阵，包括偏协方差的负数。在一个好的因子模型中除对角线上的系数较大外，远离对角线上的元素的系数应该比较小。

➤ “KMO 和巴特利特球形检验”选项：要求进行 KMO 检验和球形 Bartlett 检验。选择此项则给出对抽样充足性的 Kaiser-Meyer-Olkin 检验，检验变量间的偏相关是否很小。Bartlett 球形检验，检验相关矩阵是否单位矩阵，它表明因子模型是否不合适，也就是说数据是否适合作因子分析。

(3) “提取”子对话框的设置：单击“提取(E)…”按钮，弹出此子对话框，设置如图 10-3 所示，现对其中各选项解释如下。

① “方法”选项：下拉列表中一共有 7 种提取方法，分别说明如下。

- “主成分”法：该方法假设变量是因子的纯线性组合。第一成分有最大的方差，后续的成分方差逐个递减。主成分法是常用的获取初始因子分析结果的方法，它假设特殊因子的作用可以忽略不计。
- “未加权最小平方”法：使用未加权的最小平方方法来提取因子。未加权的最小平方方法在忽略对角线元素的情况下，最小化相关矩阵和再生矩阵差值的平方和。

- “广义最小平方”法：使用广义最小平方方法来提取因子。综合最小平方方法最小化相关矩阵和再生矩阵差值的平方和。相关性用它们值的倒数加权，以便有较高值的变量有较低的权。
- “最大似然”法：使用最大似然估计法来提取因子。最大似然估计法生成一个参数的估计，如果样本取自多维正态分布，则这个参数估计是能产生观测的相关矩阵中有最大概率的一个。相关性使用变量值的倒数进行加权，还使用了迭代算法。
- “主轴因式分解”法：使用主轴因子法来提取因子。主轴因子法使用多元相关的平方作为对公因子方差的估计值。
- “Alpha 因式分解”法：使用因子法来提取因子，最大化因子的依赖度。
- “映像因式分解”法：使用多元回归法来提取因子。它是由 Guttman 在映像理论的基础上建立起来的。变量的公共部分（称为偏映像）定义为残余变量的线性组合，而不是作为假设因子的函数。

② “分析”选项组：用于确定相关矩阵和协方差矩阵。

- “相关性矩阵”选项：使用变量的相关矩阵进行分析，当参与分析的变量的测度单位不同时，应该选择此项。
- “协方差矩阵”选项：使用变量的协方差矩阵进行分析，当参与分析的变量测度单位相同时，可以选择此项。

③ “输出”选项组：用于指定与因子提取相关的输出项。

- “未旋转因子解”选项：要求显示未经旋转的因子提取结果。此项为系统默认的输出方式。
- “碎石图”选项：要求显示按特征值大小排列的因子序号，以特征值为两个坐标轴的碎石图，可以有助于确定保留多少个因子。典型的碎石图有一个明显的拐点，在该点之前是与大因子连接的陡峭的折线，之后是与小因子相连的平缓折线。

④ “提取”选项组：选择控制提取进程和提取结果的选项。理论上因子数目与原始变量数目相等，但因子分析的目的是用少量的因子代替多个原始变量。选择提取多少个因子由本组选项决定。



图 10-3 “因子分析：提取”对话框

- “基于特征值”选项：指定提取的因子应该具有的特征值范围，在此项后面的矩形框中给出，系统默认值为 1，即要求提取那些特征值大于 1 的因子。
- “因子的固定数目”选项：指定提取公因子的数目。选择此项后，将指定的数目输入该选项后面的矩形框中，数值应该是 0 至分析变量数目之间的正整数（一般将提取数目指定在其所有特征值的方差累积贡献率达 80% 以上）。

⑤ “最大收敛性迭代次数”框：指定因子分析收敛的最大迭代次数，系统默认的最大迭代次数为 25。

（4）“因子得分”子对话框的设置：单击“得分（S）...”按钮，弹出此子对话框，设置如图 10-4 所示，现对其中各选项解释如下。

① “保存为变量”选项：如果选择此选项，则将因子得分作为一个变量保存起来。对分析结果中的每一个因子都会生成一个新变量。

② “方法”选项组：在此选项组中选定计算因子得分系数的方法。只有选择了“保存为变量”复选框后，该选项组才会被激活。有以下三种方法。

- “回归”法：其因子得分的均值为 0，方差等于估计因子得分与实际因子得分之间的多元相关的平方。
- “巴特利特”：Bartlett 法，因子得分均值为 0，超出变量范围的特殊因子平方和被最小化。
- “安德森-鲁宾”：Anderson-Rubin 法，是为了保证因子的正交性而对 Bartlett 因子得分的调整，其因子得分的均值为 0，标准差为 1，且彼此不相关。

（5）“选项”子对话框的设置：单击“选项（O）...”按钮，弹出此子对话框，设置如图 10-5 所示，现对其中各选项解释如下。



图 10-4 “因子分析：因子得分”对话框

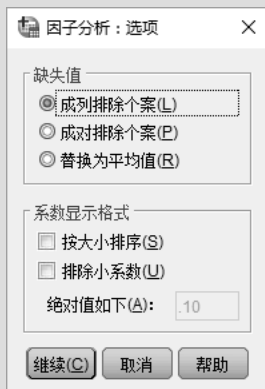


图 10-5 “因子分析：选项”对话框

① “缺失值”单选项组：有以下几种处理缺失值的方法。

- “成列排除个案”法：在分析过程中对指定的分析变量中有缺失值的观测量一律剔除，即所有带有缺失值的观测量都不参与分析。
- “成对排除个案”法：成对剔除带有缺失值的观测量。即在计算两个变量的相关系数时，只把这两个变量中带有缺失值的观测量剔除。如果一个观测量在正进行相关系数计算的变量中没有缺失值，其他变量中带有缺失值，则该观测量仍参加相关系数的计算。选择此项可以最大限度地利用得来不易的原始数据。
- “替换为平均值”法：用变量的均值代替该变量的所有缺失值。

② “系数显示格式”选项组：选择系数的显示格式，有以下几种。

- “按大小排序”法：载荷系数按其数值的大小排列并构成矩阵，使在同一因子上具有较高载荷的变量排在一起，便于得出结论。
- “排除小系数”法：不显示那些绝对值小于指定值的载荷系数。需在其后面的框中输入 0~1 之间的数作为临界值，系统默认的临界值为 0.10。选择此项可以突出载荷较大的变量，便于得出结论。

☆说明☆

- (1) 由于在 SPSS 中并没有完整的主成分分析过程，其主成分分析过程是集成在“因子分析”过程中的，但并不完善。由于主成分的得分需要对因子得分情况进行进一步计算，故不需设置“得分”子对话框，即不需保存因子得分情况，即使保存了，因子得分也不是各主成分得分的结果。
- (2) 对于提取因子的个数问题，一般遵循两个标准，其一是累积方差贡献率在 80%以上，其二是其特征值大于 1。本例之所以将要提取的因子数设置为 3，是因为通过预先分析，发现前 3 个主成分的累积方差贡献率为 86.696%，也就是说所提取的 3 个主成分可以解释总体信息的 86.696%。

以上各项设置完成后，提交系统运行。

第 4 步 因子分析的结果。

本次运行的主要结果如表 10.3、表 10.4 及图 10-6 所示，具体分析如下。

(1) 特征值和方差贡献表：表 10.3 是特征值和方差贡献表，“总计”部分为各因子对应的特征根，“方差百分比”部分为各因子的方差贡献率，“累积%”部分为累积方差贡献率。从表中可以看出，前 3 个主成分已经解释了总方差的近 86.7%，故可以选择前 3 个主成分进行分析。

表 10.3 特征值和方差贡献表

成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	6.049	40.325	40.325	6.049	40.325	40.325
2	5.813	38.755	79.080	5.813	38.755	79.080
3	1.142	7.616	86.696	1.142	7.616	86.696
4	.876	5.842	92.538			
5	.599	3.996	96.534			
6	.326	2.174	98.709			
7	.119	.796	99.505			
8	.041	.272	99.776			
9	.018	.121	99.897			
10	.010	.063	99.961			
11	.004	.027	99.988			
12	.001	.009	99.997			
13	.000	.002	99.999			
14	.000	.001	100.000			
15	4.080E-7	2.720E-6	100.000			

提取方法：主成分分析法。

(2) 主成分的碎石图：图 10-6 是主成分的碎石图，结合特征根曲线的拐点及特征根值，从图上可看出，前 3 个主成分的折线坡度较陡，而后面就逐渐趋于平缓，该图从另一个侧面说明了取前三个主成分为宜。

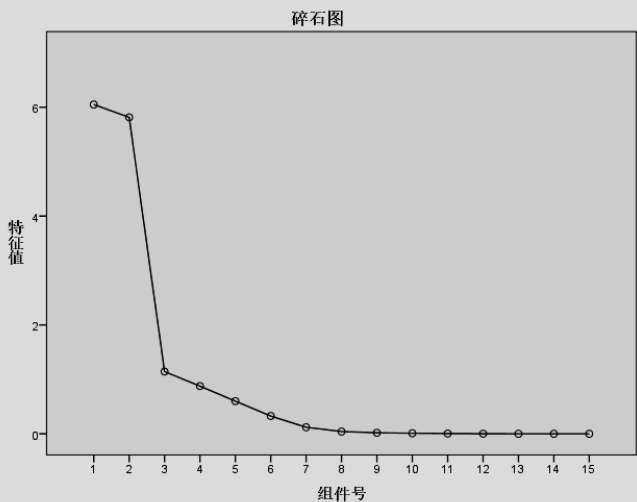


图 10-6 各成分的碎石图

(3) 旋转前的因子载荷矩阵，如表 10.4 所示。

表 10.4 旋转前的因子载荷矩阵

	成分		
	1	2	3
x1	.407	.805	.268
x2	.596	-.727	.209
x3	-.147	.016	.821
x4	.895	-.333	-.181
x5	.614	.763	.028
x6	.826	-.124	-.281
x7	.273	-.627	.184
x8	.636	.703	.041
x9	.619	.703	.008
x10	.552	.766	.196
x11	.654	-.691	.172
x12	.666	-.685	.166
x13	.863	-.191	-.297
x14	.728	-.632	.144
x15	.579	.760	.005

提取方法：主成分分析法。

a. 提取了3个成分。

☆说明☆

◆ 式 (10.7) 中的 $e'_i = (e_{1i}, e_{2i}, \dots, e_{pi})$ 是标准化正交向量，并不是 SPSS 输出“因子载荷矩阵”中的系数。而“因子载荷矩阵”中各分量的系数为单位特征向量乘以相应的特征值的平方根的结果，其公式为 $e_{ij} = a_{ij} / \sqrt{\lambda_i}$ 。故需进一步利用因子分析的结果进行主成分分析。

第 5 步 利用因子分析的结果进行主成分分析。

表 10.4 是旋转前的因子载荷矩阵，并不是主成分分析中所需要的标准化正交向量，要得到标准化正交向量还需进行如下运算。

- (1) 将表 10.4 因子载荷矩阵中的数据输入 SPSS 数据编辑窗口中, 将 3 个变量名分别命名为 a_1 , a_2 和 a_3 。
- (2) 用公式 $e_{ij} = a_{ij} / \sqrt{\lambda_i}$ 计算出标准化特征向量。步骤: 打开“转换→计算变量”, 计算过程如图 10-7 所示, 其中 5.813 为第 2 个特征值。对 t_1 , t_2 和 t_3 须分别进行计算。
- (3) 计算结束后得到的特征向量矩阵如表 10.5 所示。



图 10-7 标准化正交向量的计算图示

表 10.5 标准化正交特征向量矩阵

变量	t_1	t_2	t_3
x_1	0.17	0.33	0.25
x_2	0.24	0.30	0.20
x_3	0.06	0.01	0.77
x_4	0.36	0.14	0.17
x_5	0.25	0.32	0.03
x_6	0.34	0.05	0.26
x_7	0.11	0.26	0.17
x_8	0.26	0.29	0.04
x_9	0.25	0.29	0.01
x_{10}	0.22	0.32	0.18
x_{11}	0.27	0.29	0.16
x_{12}	0.27	0.28	0.16
x_{13}	0.35	0.08	0.28
x_{14}	0.30	0.26	0.13
x_{15}	0.24	0.32	0.00

- (4) 对原始数据变量进行标准化。由于是以相关系数矩阵为出发点进行因子分析的, 所以主成分分析表达式中的变量应该是经过标准化的数据。标准化变量通过“分析→描述统计→描述”命令实现, 在“描述性”对话框中将需要标准化的 $x_1 \sim x_{15}$ 选入“变量”对话框后, 选中底部的“将标准化得分另存为变量”, 就会在数据窗口中增加 $Zx_1 \sim Zx_{15}$ 共 15 个变量, 它们分别是 $x_1 \sim x_{15}$ 的标准化变量。

(5) 计算主成分: 主成分的计算公式为式 (10.7), 其表达式为 $y = Zx * t$, 其中 Zx 为变量标准化后的矩阵, t 为如表 10.5 所示的标准化的正交特征向量矩阵, 矩阵的乘法可在 Excel 中进行, 亦可在 MATLAB 中进行。再通过表 10.3 所示各主成分分析的方差百分比(第 1 主成分占 40.325%, 第 2 主成分占 38.755%, 第 3 主成分占 7.616%), 计算出综合得分函数, 其公式为 $y_{\text{综}} = 0.40325y_1 + 0.38755y_2 + 0.07616y_3$ 。则各主成分及综合得分情况如表 10.6 所示。通过综合得分的高低 ($y_{\text{综}}$) 可知各国参与国际化水平的高低, 其中美国最高, 印度最低。

表 10.6 主成分及综合得分情况

编号	国家	y_1	y_2	y_3	$y_{\text{综}}$
1	中国	-2.19	0.07	3.01	-0.63
2	印度	-2.56	-0.11	-0.46	-1.11
3	日本	0.45	1.85	-0.27	0.88
4	韩国	-1.69	-0.46	-0.27	-0.88
5	新加坡	5.28	-6.26	1.19	-0.20
6	美国	3.30	6.07	1.46	3.80
7	加拿大	-0.43	-0.47	-0.31	-0.38
8	巴西	-1.91	-0.06	-0.43	-0.83
9	墨西哥	-1.68	-0.68	0.03	-0.94
10	英国	4.46	0.98	-1.75	2.05
11	法国	0.87	0.46	-0.52	0.49
12	德国	1.40	1.34	-0.26	1.06
13	意大利	-0.61	0.10	-0.54	-0.25
14	俄罗斯	-2.35	-0.20	-0.30	-1.05
15	澳大利亚	-1.36	-0.92	-0.30	-0.93
16	新西兰	-0.99	-1.73	-0.28	-1.09

10.3 因子分析

10.3.1 基本概念及统计原理

1. 统计原理

因子分析的出发点是用较少的相互独立的因子变量来代替原来变量的大部分信息，可以用下面的数学模型来表示：

$$\begin{cases} x_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m \\ x_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m \\ \vdots \\ x_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m \end{cases} \quad (10.8)$$

式中， x_1, x_2, \cdots, x_p 为 p 个原有变量，是均值为零、标准差为 1 的标准化变量， F_1, F_2, \cdots, F_m 为 m 个因子变量， m 小于 p ，表示成矩阵形式：

$$X = AF + a\varepsilon \quad (10.9)$$

式中， F 为因子变量或公共因子，可以理解为高维空间中互相垂直的 m 个坐标轴； A 为因子载荷矩阵，是第 i 个原有变量在第 j 个因子变量上的负荷；如果把变量 x_i 看成 m 维因子空间中的一个向量，则 a_{ij} 为 x_i 在坐标轴 F_j 上的投影，相当于多元回归中的标准回归系数； ε 为特殊因子，表示了原有变量不能被因子变量所解释的部分，相当于多元回归分析中的残差部分。

2. 分析步骤

因子分析有两个核心问题：一是如何构造因子变量；二是如何对因子变量进行命名解释。因子分析有以下 5 个基本步骤。

第 1 步 将原始数据进行标准化。

进行因子分析是在标准化数据的基础上进行的，所以须将原始数据标准化。

第 2 步 确定待分析的原有若干变量是否适合于因子分析。

进行因子分析要求原有变量之间存在较强的相关性，如果没有较强的相关关系，则无法从中综合出能反映某些变量共同特征的少数公共因子变量来。

SPSS 提供的几种检验变量是否适合于因子分析的方法有：

(1) 巴特利特球形检验 (Bartlett Test of Sphericity)。以变量的相关系数矩阵为出发点，其原假设是相关系数矩阵为一个单位阵，其统计量是根据相关系数矩阵的行列式得到的，如果该值较大，且其相伴概率小于显著性水平，则应拒绝原假设，说明原始矩阵不可能是单位阵，即原变量之间存在相关性，适宜作因子分析；反之，不宜作因子分析。

(2) 反映像相关矩阵检验 (Anti-image correlation matrix)。以变量的偏相关系数矩阵为出发点，将偏相关系数矩阵的每个元素取反，得到反映像相关矩阵。偏相关系数是在控制了其他变量对两变量影响的条件下计算出来的相关系数，如果变量之间存在较多的重叠影响，那么偏相关系数就会较小。因此，如果反映像相关矩阵中有些元素的绝对值比较大，则说明这些变量不适合作因子分析。

(3) KMO (Kaiser-Meyer Olkin) 检验。KMO 值越接近于 1，则所有变量之间的简单相关系数平方和远大于偏相关系数平方和，因此越适合于作因子分析；KMO 值越小，越不适合于作因子分析。

第 3 步 构造因子变量。

建立变量的相关系数矩阵 R ，求 R 的特征根及相应的单位特征向量，根据累积贡献率的要求（或特征值大小的要求），取前 m 个特征根及相应的特征向量，写出因子载荷矩阵 A 。

第 4 步 利用旋转使得因子变量更具有可解释性。

将原有变量综合为少数几个因子后，如果因子的实际含义不清，则极不利于进一步分析。一般需利用旋转方法使提出的因子含义更加清晰，使因子具有命名可解释性。

第 5 步 计算因子变量的得分。

因子变量确定后，对每个样本数据，我们希望得到它们在不同因子上的具体数值，这些数值就是因子得分，它和原变量的得分相对应。有了因子得分，我们在以后的研究中就可以针对维数少的因子得分来进行。

计算因子得分的模型：

$$F_j = \beta_{j1}X_1 + \cdots + \beta_{jp}X_p \qquad j=1, 2, \cdots, m \qquad (10.10)$$

估计因子得分的方法很多，如加权最小二乘法、回归法等。

10.3.2 因子分析 SPSS 实例分析

【例 10-2】 为了研究几个省市的科技创新力问题，现取了 8 个省市某年的 15 个科技指标数据，试分析一个省的科技创新能力主要受哪些潜在因素的影响？（数据来源：屠文娟，《基于因子分析法的江苏省科技原创力评价与提升》，科技管理研究；参见数据文件：data10-2.sav。）

表 10.7 8 个省市某年的科技指标数据

省市	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
北京	229	80.26	48.5	24.49	3.55	5.55	10.23	44774.45	25.02	24.1	779.24	226.01	34.42	3183.29	2.12
天津	87	67.48	36.82	14.08	2.62	1.96	4.49	35451.77	33.59	21.38	410.34	73.15	25.06	495.78	1.82
辽宁	44	65.69	35.94	8.34	2.32	1.56	2.45	18974.2	11.29	5.57	263.35	22.32	15.21	204.98	1.78
上海	104	74.06	35.98	17.84	4.78	2.28	4.8	51485.83	39.72	19.08	654.31	112.32	15.85	1303.32	2
江苏	50	60.79	34.07	6.8	2.13	1.47	3.17	24489.18	43.13	17.99	206.68	16.6	9.14	134.89	1.41
浙江	53	63.48	31.08	5.42	3.95	1.22	1.83	27435.38	7.94	7.63	257.65	22.66	5.82	79.01	1.72
山东	30	64.59	33.22	4.44	1.81	1.05	1.59	20022.57	9.17	5.69	117.73	9.76	8.41	106.36	1.34
广东	35	69.64	37.27	5.81	3.66	1.09	2.18	24327.32	35.67	24.99	117.51	20.4	5.08	122.33	1.47

表 10.8 对表 10.7 指标意义的解释

指标	指标意义
x_1	每万人口科技活动人员数（人/万人）
x_2	从事科技活动人员中科学家工程师所占比重（%）
x_3	R&D 人员占科技活动人员的比重（%）
x_4	大专以上学历人口数占总人口数的比例（%）
x_5	地方财政科技拨款占地方财政支出的比重（%）
x_6	R&D 经费总量占 GDP 比重（%）
x_7	R&D 经费中基础研究经费所占比例（%）
x_8	人均 GDP（元/人）
x_9	高技术产品出口额占商品出口额的比重（%）
x_{10}	规模以上产业增加值中高技术产业份额（%）
x_{11}	万名科技人员被国际三大检索工具收录的论文数（篇/万人）
x_{12}	每百万人口发明专利的授权量（件/百万人）
x_{13}	发明专利申请授权量占专利申请授权量的比重（%）
x_{14}	万人技术市场成交合同金额（万元/万人）
x_{15}	财政性教育经费支出占 GDP 比重（%）

第 1 步 分析。

如题所述，要分析一个省的科技创新能力受哪些潜在因素的影响，可用因子分析法进行分析。

第 2 步 数据组织。

建立 $x_1 \sim x_{15}$ 共 15 个数据变量和一个“省市”字符型变量，将北京、天津等 8 个省市作为个案数据输入并保存。

第 3 步 因子分析设置。

(1) 按“分析→降维→因子”打开“因子分析”对话框，将 $x_1 \sim x_{15}$ 这 15 个变量移入“变量”对话框中，表示对这 15 个变量数据进行因子分析。

(2) “提取”对话框的设置：单击“提取(E)…”按钮，弹出此子对话框，设置如图 10-8 所示。各项的意义在图 10-3 中已解释，这里不再赘述。与图 10-3 设置不同的是在特征值的选取上，选择了“基于特征值”单选项，并在“特征值大于”栏中设置为 1，表示取特征值大于 1 的公共因子。

(3) “旋转”对话框的设置：单击“旋转(T)…”按钮，弹出此子对话框，设置如图 10-9 所示，现对其中各项解释如下。



图 10-8 “因子分析：提取”对话框



图 10-9 “因子分析：旋转”对话框

① “方法”单选项组：选择旋转的方法。主要有以下几种。

- “无”：不进行因子旋转。
- “最大方差法”：也称正交旋转法。它将每一个有最大负荷的因子的变量数最小化，因此可以简化对因子的解释。
- “直接斜交法”：直接斜交旋转，选定该项，可以在下面的矩形框中输入 Delta 值，该值介于 0 和 1 之间，0 表示产生最高的相关系数。
- “四次幂极大法”：四分最大正交旋转，该旋转法使每个变量中需要解释的因子数最少。
- “等量最大法”：平均正交旋转，是“最大方差法”和“最大四次方值法”的结合，表示全体旋转，对变量和因子均旋转。
- “最优斜交法”：它比直接斜交旋转更快，因此适用于大数据的因子分析。选择该法，在其下的 Kappa 栏内输入控制斜交旋转的参数，系统默认为 4。

② “输出”选项组：选择输出哪些结果。有以下两个选项。

- “旋转后的解”选项：当在“方法”栏中选择了一种旋转方法后，此选项才被激活。对

正交旋转，输出旋转模型矩阵、因子转换矩阵。对斜交旋转，则输出模型、结构和因子相关矩阵。

➤ “载荷图”选项：选择此项，则输出前两个因子的二维载荷图，或前三个因子的三维载荷图，如果仅提取一个公因子，则不输出载荷图。

(4) “得分”对话框的设置：单击“得分(S)…”按钮，弹出此子对话框，选择“保存为变量”，即将因子得分保存下来。并选择“方法”下的“回归”方法，表示以回归方法确定因子得分。

以上各项设置完成后提交系统运行。

第 4 步 主要结果及分析。

因子分析的运行结果如表 10.9~表 10.12 所示，具体分析如下。

(1) 特征值与方差贡献表：表 10.9 是特征值与方差贡献表，可以看出前 3 个特征值大于 1，同时这 3 个公共因子的方差贡献率占了 93.924%，说明提取这 3 个公共因子可以解释原变量的绝大部分信息。

表 10.9 特征值与方差贡献表

成分	初始特征值			提取载荷平方和			旋转载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	11.135	74.237	74.237	11.135	74.237	74.237	9.042	60.280	60.280
2	1.706	11.371	85.608	1.706	11.371	85.608	2.926	19.507	79.787
3	1.247	8.316	93.924	1.247	8.316	93.924	2.120	14.137	93.924
4	.508	3.386	97.310						
5	.205	1.365	98.675						
6	.125	.832	99.507						
7	.074	.493	100.000						
8	1.007E-15	6.711E-15	100.000						
9	2.374E-16	1.582E-15	100.000						
10	8.285E-17	5.524E-16	100.000						
11	-5.178E-17	-3.452E-16	100.000						
12	-1.360E-16	-9.064E-16	100.000						
13	-5.414E-16	-3.609E-15	100.000						
14	-6.354E-16	-4.236E-15	100.000						
15	-1.388E-15	-9.256E-15	100.000						

提取方法：主成分分析法。

(2) 旋转前的因子载荷矩阵：表 10.10 是旋转前的因子载荷矩阵，表的底部表明使用的是主成分分析法，3 个主成分被抽取出来。

(3) 旋转后的因子载荷矩阵：表 10.11 是按照前面设定的“方差极大法”对因子载荷矩阵旋转的结果。在表 10.10 所示未经旋转的载荷矩阵中，因子变量在许多变量上均有较高的载荷，从旋转后的因子载荷矩阵可以看出，因子 1 在 1、3、4、6、7、12、13、14 上有较大载荷，反映科技投入与产出情况，可以命名为创新水平因子；因子 2 在指标 5、8、15 上有较大载荷，反映地区经济发展及财政科教投入水平，可以命名为创新环境因子；因子 3 在指标 9 和指标 10 上有较大载荷，可以命名为高技术产业发展因子。

(4) 因子转换矩阵表：表 10.12 是因子转换矩阵表，表明因子提取的方法是主成分分析法，旋转的方法是方差极大法。

表 10.10 旋转前的因子载荷矩阵

	成分		
	1	2	3
x1	.973	-.158	.052
x2	.919	.036	-.090
x3	.883	-.161	.334
x4	.985	-.004	-.022
x5	.482	.497	-.664
x6	.947	-.242	.131
x7	.972	-.108	.178
x8	.849	.340	-.301
x9	.300	.834	.386
x10	.611	.637	.399
x11	.955	-.001	-.211
x12	.992	-.091	-.001
x13	.876	-.282	.205
x14	.968	-.156	.032
x15	.859	-.092	-.385

提取方法：主成分分析法。
a. 提取了 3 个成分。

表 10.11 旋转后的因子载荷矩阵

	成分		
	1	2	3
x1	.936	.286	.130
x2	.776	.459	.202
x3	.924	.016	.251
x4	.867	.413	.221
x5	.068	.940	.180
x6	.966	.177	.095
x7	.944	.202	.235
x8	.541	.726	.327
x9	.018	.137	.956
x10	.377	.172	.876
x11	.794	.558	.118
x12	.913	.365	.161
x13	.937	.071	.084
x14	.926	.301	.119
x15	.705	.626	-.069

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。
a. 旋转在 5 次迭代后已收敛。

(5) 因子得分及综合因子得分情况：各因子的得分已保存在数据文件中。综合因子得分为 $F = 0.6028F_1 + 0.19507F_2 + 0.14137F_3$ 。可通过“转换→计算变量”进行计算并排序，其结果如表 10.13 所示。

表 10.12 因子转换矩阵

成分	1	2	3
1	.884	.403	.239
2	-.405	.400	.822
3	.236	-.823	.517

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。

表 10.13 因子得分及综合因子得分

省市	F ₁	F ₂	F ₃	F	综合排序
山东	-0.344	-1.001	-0.945	-0.536	8
浙江	-0.791	0.905	-1.223	-0.473	7
江苏	-0.488	-1.024	1.073	-0.342	6
广东	-0.791	-0.104	1.202	-0.327	5
辽宁	-0.002	-0.500	-1.206	-0.269	4
天津	0.248	-0.275	0.572	0.177	3
上海	-0.136	1.947	0.481	0.366	2
北京	2.305	0.053	0.045	1.406	1

10.4 典型案例

10.4.1 医院工作质量评价分析

为了评价医院的工作质量，某研究者收集了某医院某三年与医院质量相关的 9 个指标，分别是 x_1 ：门诊人次（万人）， x_2 ：出院人数， x_3 ：病床利用率（%）， x_4 ：病床周转次数， x_5 ：平均住

院天数, x_6 : 治愈好转率 (%), x_7 : 病死率 (%), x_8 : 诊断符合率 (%), x_9 : 抢救成功率 (%)。具体数据如表 10.14 所示。试分析医院的工作质量究竟受哪些主要因素的影响。(数据来源: 赖国毅等,《SPSS 17.0 中文版常用功能与应用》, 电子工业出版社; 参见数据文件: data10-3.sav。)

表 10.14 某医院某三年医疗工作质量指标数据

年月	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1-01	4.34	389	99.06	1.23	25.46	93.15	3.56	97.51	61.66
1-02	3.45	271	88.28	0.85	23.55	94.31	2.44	97.94	73.33
1-03	4.38	385	103.97	1.21	26.54	92.53	4.02	98.48	76.79
1-04	4.18	377	99.48	1.19	26.89	93.86	2.92	99.41	63.16
1-05	4.32	378	102.01	1.19	27.63	93.18	1.99	99.71	80
1-06	4.13	349	97.55	1.1	27.34	90.63	4.38	99.03	63.16
1-07	4.57	361	91.66	1.14	24.89	90.6	2.73	99.69	73.53
1-08	4.31	209	62.18	0.52	31.74	91.67	3.65	99.48	61.11
1-09	4.06	425	83.27	0.93	26.56	93.81	3.09	99.48	70.73
1-10	4.43	458	92.39	0.95	24.26	91.12	4.21	99.76	79.07
1-11	4.13	496	95.43	1.03	28.75	93.43	3.5	99.1	80.49
1-12	4.1	514	92.99	1.07	26.31	93.24	4.22	100	78.95
2-01	4.11	490	80.9	0.97	26.9	93.68	4.97	99.77	80.53
2-02	3.53	344	79.66	0.68	31.87	94.77	3.59	100	81.97
2-03	4.16	508	90.98	1.01	29.43	95.75	2.77	98.72	62.86
2-04	4.17	545	92.98	1.08	26.92	94.89	3.14	99.41	82.35
2-05	4.16	507	95.1	1.01	25.82	94.41	2.8	99.35	60.61
2-06	4.86	540	93.17	1.07	27.59	93.47	2.77	99.8	70.21
2-07	5.06	552	84.38	1.1	27.56	95.15	3.1	98.63	69.23
2-08	4.03	453	72.69	0.9	26.03	91.94	4.5	99.05	60.42
2-09	4.15	529	86.53	1.05	22.4	91.52	3.84	98.58	68.42
2-10	3.94	515	91.01	1.02	25.44	94.88	2.56	99.36	73.91
2-11	4.12	552	89.14	1.1	25.7	92.65	3.87	95.52	66.67
2-12	4.42	597	90.18	1.18	26.94	93.03	3.76	99.28	73.81
3-01	3.05	437	78.81	0.87	23.05	94.46	4.03	96.22	87.1
3-02	3.94	477	87.34	0.95	26.78	91.78	4.57	94.28	87.34
3-03	4.14	638	88.57	1.27	26.53	95.16	1.67	94.5	91.67
3-04	3.87	583	89.82	1.16	22.66	93.43	3.55	94.49	89.07
3-05	4.08	552	90.19	1.1	22.53	90.36	3.47	97.88	87.14
3-06	4.14	551	90.81	1.09	23.06	91.65	2.47	97.72	87.13
3-07	4.04	574	81.36	1.14	26.65	93.74	1.61	98.2	93.02
3-08	3.93	515	76.87	1.02	23.88	93.82	3.09	95.46	88.37
3-09	3.9	555	80.58	1.1	23.08	94.38	2.06	96.82	91.79
3-10	3.62	554	87.21	1.1	22.5	92.43	3.22	97.16	87.77
3-11	3.75	586	90.31	1.12	23.73	92.47	2.07	97.14	93.89
3-12	3.77	627	86.47	1.24	23.22	91.17	3.4	98.98	89.8

案例分析：研究者一共收集了 9 个指标，如果直接用这 9 个指标进行评价，由于指标太多，一是会给数据收集和评价的计算量带来麻烦，二是由于部分指标之间具有较强的相关性，这也会对评价结果的正确性产生影响。所以，有必要对这 9 个指标进行“降维”处理，可采用主成分分析或因子分析法进行，找出主要成分或公共因子后进行评价。

10.4.2 各省、市、自治区城市市政设施建设状况分析

为了了解我国各省、市、自治区城市市政设施建设状况，选取了 2015 年全国所有省、市、自治区的 6 个指标进行统计分析，分别是 x_1 ：年末实有道路长度（万公里）， x_2 ：年末实有道路面积（万平方米）， x_3 ：城市桥梁（座）， x_4 ：城市排水管道长度（万公里）， x_5 ：城市污水日处理能力（万立方米）， x_6 ：城市路灯（盏）。具体数据如表 10.15 所示，请分析各省、市、自治区的市政建设情况。（数据来源：《中国统计年鉴》，2016 年；参见数据文件：data10-4.sav。）

表 10.15 各省、市、自治区市政设施指标

省、市、自治区	x_1	x_2	x_3	x_4	x_5	x_6
北 京	0.81	14302	2271	1.55	461.7	239840
天 津	0.76	14019	953	1.95	285.9	350682
河 北	1.34	31570	1410	1.70	564.4	660167
山 西	0.73	15039	657	0.79	244.9	584263
内蒙古	0.93	19793	349	1.25	212.0	721653
辽 宁	1.69	30585	1637	1.71	831.4	1580767
吉 林	0.89	17010	776	1.03	334.5	502560
黑龙江	1.24	18651	1065	1.03	736.9	635808
上 海	0.50	10317	2524	1.69	785.0	532032
江 苏	4.07	75052	14630	7.00	1673.2	3391262
浙 江	2.05	39293	10283	3.82	882.1	1458067
安 徽	1.34	31010	1620	2.44	646.2	848632
福 建	0.84	16303	1816	1.33	380.8	718378
江 西	0.82	17436	827	1.20	272.9	649981
山 东	4.04	80847	5212	5.22	1004.2	1864921
河 南	1.23	29915	1374	2.05	649.8	863653
湖 北	1.79	31852	1952	2.30	656.0	530996
湖 南	1.14	21333	827	1.32	586.7	182019
广 东	3.89	70003	6259	5.36	1974.6	2301867
广 西	0.82	17003	832	1.06	689.6	627042
海 南	0.24	5070	226	0.38	93.4	181211
重 庆	0.77	16128	1433	1.30	273.8	511560
四 川	1.34	27937	2081	2.25	565.9	984428
贵 州	0.35	7201	683	0.59	144.0	421684
云 南	0.62	12690	693	1.15	245.2	473939
西 藏	0.10	1891	18	0.14	7.4	59648
陕 西	0.65	14602	743	0.80	337.5	638418
甘 肃	0.45	9650	540	0.56	160.2	296520
青 海	0.10	1970	144	0.17	41.9	120010
宁 夏	0.21	6376	188	0.16	82.0	266358
新 疆	0.74	12827	489	0.65	241.3	562837

案例分析：如果直接使用原始数据对这 31 个省、市、自治区的市政设施进行评价，不但使问题变得烦琐，而且也不便于分析这几个指标对其他相关指标的影响。为此，可以进行因子分析（或主成分分析），提取出几个公因子（或主成分），并在提取的公因子（或主成分）的基础上进行回归分析，最后根据综合评价高低对各省、市、自治区的市政建设情况进行排序。

10.4.3 大学生的价值观分析

为了研究大学生的价值观，某研究人员抽样调查了 20 名大学生关于价值观的 9 项检验结果。包括：合作性 (x_1)，对分配的看法 (x_2)，行为出发点 (x_3)，工作投入程度 (x_4)，对发展机会的看法 (x_5)，对社会地位的看法 (x_6)，权力距离 (x_7)，对职位升迁的态度 (x_8)，领导风格的偏好 (x_9)。分值区间为[1, 20]，具体数据如表 10.16 所示。请分析影响大学生价值观的因素。（数据来源：武松等，《SPSS 统计分析大全》，清华大学出版社；参见数据文件：data10-5.sav。）

表 10.16 20 名大学生价值观检验数据

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	16	16	13	18	16	17	15	16	16
2	18	19	15	16	18	18	18	17	19
3	17	17	17	14	17	18	16	16	16
4	17	17	17	16	19	18	19	20	19
5	16	15	16	16	18	18	15	16	16
6	20	17	16	17	18	18	17	19	18
7	18	16	16	20	15	16	19	14	17
8	20	18	18	17	18	19	18	19	18
9	14	16	15	19	19	19	18	19	14
10	19	19	20	14	18	20	19	17	20
11	19	19	14	14	16	17	16	17	18
12	15	15	18	16	18	18	19	17	18
13	16	17	15	17	15	18	16	14	13
14	17	14	12	14	14	18	15	15	13
15	14	16	14	15	16	16	17	16	17
16	10	11	13	18	17	20	17	16	20
17	16	17	15	16	14	16	14	15	17
18	15	16	15	17	16	16	16	15	16
19	16	19	18	15	17	12	19	18	18
20	16	16	13	18	16	17	15	16	16

10.5 思考与练习

1. 主成分分析和因子分析常用来解决什么问题？它们的目标是什么？
2. 在用 SPSS 进行主成分分析时应注意哪些问题？
3. 已知影响粮食产量的指标有 x_1 ：农村劳动力（万人）， x_2 ：播种面积（万亩）， x_3 ：有效灌溉面积（万亩）， x_4 ：化肥施用量（万吨）， x_5 ：大牲畜存栏数（万头）， x_6 ：生猪存栏数（万口）6 个指标，今调查某省 10 个产粮区的数据，如表 10.17 所示，试分别对其进行主成分分析和因子分析。（数据来源：郝黎仁，《SPSS 实用统计分析》，中国水利水电出版社；参见数据文件：data10-6.sav。）

表 10.17 某省 10 个产粮区的指标数据

编号	x_1	x_2	x_3	x_4	x_5	x_6
1	74.94	498.19	9.25	119.47	39.26	48.13
2	77	129.64	6.7	88.08	23.91	35.84
3	81.82	201.11	12.9	148.26	42.53	39.03

续表						
4	78.42	203.45	14.93	158.87	44.64	56.56
5	81.44	619.48	6.65	128.55	61.17	85.18
6	84.71	467.02	6.17	111.99	56.54	62.94
7	77.33	508.17	6.32	126.86	48.22	43.16
8	84.65	613.55	8.25	187.19	54.61	40.33
9	85.55	202.27	4.49	88.5	30.79	21.3
10	73.55	319.48	4.13	107.61	23.08	37.23

4. 某研究机构从载文量、基金论文比等 8 项指标对 15 所高校学报的学术影响力进行了研究，具体数据如表 10.18 所示，现要求从中提取出能够体现期刊学术影响水平的潜在因素，即公共因子。（数据来源：李昕等，《SPSS22.0 统计分析—从入门到精通》，电子工业出版社；参见数据文件：data10-7.sav。）

表 10.18 15 所高校学报的 8 项学术影响力指标

学报编号	载文量	基金论文比	被引期刊数	总被引频次	影响因子	即年指标	被引半衰期	Web 即年下载率
1	258	1.8	586	1158	0.529	0.039	4.6	45.4
2	279	0.72	625	1052	0.537	0.054	4.7	36.9
3	153	0.78	279	407	0.365	0.033	4.8	30.8
4	450	1	597	1461	0.593	0.058	4.5	49.5
5	226	0.82	727	1503	0.756	0.088	5	47.6
6	82	0.65	139	155	0.172	0	5.5	27.4
7	128	0.23	123	148	0.249	0.078	3	28.3
8	62	0.44	159	211	0.272	0	4.2	17.8
9	155	0.59	309	453	0.324	0.019	4.5	31.6
10	453	0.77	931	2113	0.502	0.06	5	41.2
11	160	0.99	573	1026	0.518	0.075	6.1	29.7
12	334	0.85	762	1646	0.726	0.057	4.8	40.7
13	290	0.66	480	893	0.421	0.048	4.9	31.4
14	130	0.6	302	982	0.657	0.062	5.6	30.2
15	181	0.55	367	726	0.395	0.022	5.1	37.6

第 11 章 时间序列分析

时间序列分析是多元统计分析的一项重要内容，时间序列是指按时间顺序取得的观测资料的集合。很多数据以时间序列形式呈现，如货运码头的逐月吞吐量、公路交通次数周度报告、城市空气污染物的日均值序列、医院每日门诊接诊人数序列、地区工业总产值的年度数据序列、逐年人口统计资料等。时间序列区别于普通资料的本质特征是相邻观测值之间的依赖性 or 相关性，这种特征使得时间序列资料的统计分析方法区别于一般数据的统计分析方法。事实上，有关时间序列分析的特殊技巧，几乎都是基于对自相关性处理的技巧。分析时间序列数据，可以从运动的角度认识事物的本质，如几个时间序列之间的差别、一个较长时间序列的周期性，或对未来情况进行预测。

本章将对时间序列数据的预处理、指数平滑法、自回归综合移动平均（ARIMA）法及季节分解法进行介绍。

11.1 时间序列的建立和平稳化

在对数据用时间序列模型进行拟合处理前，应先对数据进行必要的观察和预处理，直到它平稳后再用这些过程对其进行分析。（判断序列是否平稳可以看它的均值和方差是否不再随时间的变化而变化，自相关系数是否只与时间间隔有关而与所处的时间无关。）因此，根据对数据建模前预处理工作的先后顺序，将它分为三个步骤：首先，对有缺失值的数据进行补齐；其次，将数据资料定义为相应的时间序列；最后，对时间序列数据的平稳性进行计算观察。如果数据文件中存在一个变量，其值是按某一时间间隔采集的，要进行时间序列分析，还需要有一个表明采集时间的日期变量。

11.1.1 填补缺失值

时间序列分析中的缺失值不能采用通常删除的办法来解决，因为这样会导致原有时间序列周期性的破坏，而无法得到正确的分析结果。

【例 11-1】 某企业从 1998 年 1 月到 2010 年 12 月的销售数据（单位：百万元）按月记录，共 156 个观测值，如表 11.1 所示。（参见数据文件：data11-1.sav.）

表 11.1 某企业 1998~2010 年的月销售数据

	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
1998	39.01	44.24	39.5	38.25	38.04	44.15	43.52	46.65	42.3	41.87	39.52	35.18
1999	44.25	43.42	38.9	39.81	39.25	40.96	42.71	45.06	42.49	43.14	43.04	35.31
2000	40.09	46.62	39.65	37.19	39.08	43.7	44.49	49.65	44.46	44.69	42.01	38.17
2001	44.72	47.35	40.44	42.56	44.1	45.65	45.48	50.65	47.37	48.99	46.14	42.03
2002	48.94	52.56	41.4	44.72	41.85	50.92	51.47	55.87	49.91	51.23	50.44	44.63
2003	50.77	53.79	47.13	46.29	50.47	53	53.55	53.49	51.34	55.42	50.94	47.73
2004	55.45	59.83	50.91	50.55	51.18	56.49	60.57	61.63	56.86	57.02	56.34	50.29

续表

	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
2005	58.36	62.29	55.88	57.32	56.39	63.28	63.69	65.9	64.45	65.09	60.57	58.17
2006	67.68	68.83	62.67	63.16	61.96	68.44	72.08	68.76	70.98	71.81	68.36	62.73
2007	69.59	75	70.08	68.14	68.97	77.88	77.4	78.73	78.4	80.69	72.46	73.21
2008	78.21	82.27	77.18	75.03	77.68	80.97	85.07	88.33	83.34	85.7	80.5	77.18
2009	86.72	90.2	85.22	80.38	82.78	90.55	92.07	92.74	91.77	92.96	89.69	83.62
2010	94.28	98.89	91.09	93.84	94.17	103.06	102.29	102.31	100.15	101.03	101.27	97.94

第 1 步 观测是否有缺失值。

我们发现并没有缺失数据，为了练习，请读者自己删除几个数据进行练习。

第 2 步 数据组织。

将数据组织成 3 列，第 1 列是“年份”，第 2 列是“月份”，第 3 列是“销售额”，输入数据并保存。

第 3 步 缺失数据填补的设置。

按“转换→替换缺失值”打开“替换缺失值”对话框，将“销售额”变量移入“新变量”对话框中，如图 11-1 所示。

替换缺失值后，会增加一个新变量“销售额_1”。缺失值的替换根据不同的方法会得到不同的结果，SPSS 一共提供了 5 种方法：“序列平均值”、“邻近点的平均值”、“邻近点的中位数”、“线性插值”、“邻近点的线性趋势”。如果选择了“邻近点的平均值”和“邻近点的中位数”两种方法，还需在下面填上邻近点的跨度值。默认方法为“序列平均值”，如要换成其他方法，则将相应的方法选择后，再单击“变化量(H)”按钮。

具体过程和结果请读者自己练习。



图 11-1 “替换缺失值”对话框

11.1.2 定义日期变量

定义日期模块可以产生周期性的时间序列日期变量。使用“定义日期”对话框定义日期变量，需要在数据窗口读入一个按某种时间顺序排列的数据文件，数据文件中的变量名不能与系统默认的时间变量名重复，否则系统建立的日期变量会覆盖同名变量。系统默认的变量名有：年份、年份、季度、年份、月份、年份、季度、月份、日、星期、日、日、小时等。

【例 11-2】 沿用例 11-1 的数据文件，试定义日期变量。

第 1 步 定义日期变量的设置。

按“数据→定义日期和时间”顺序打开“定义日期”对话框，设置如图 11-2 所示。由于本例中的数据是从 1998 年开始的，每个月均有一个销售额数据，所以时间为“年份、月份”类型，且起始年份为 1998 年，起始月份为 1 月。

第 2 步 结果及分析。

运行完成后，在数据文件中增加了 3 个变量，分别是“YEAR_”、“MONTH_”及“DATE_”，具体如图 11-3 所示。

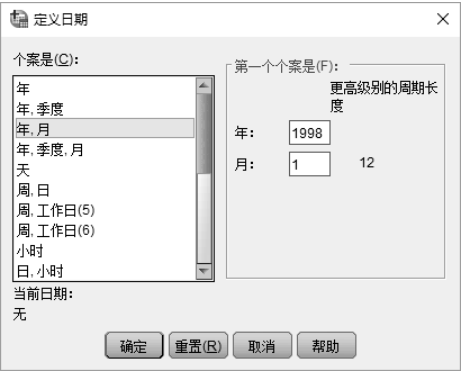


图 11-2 “定义日期”对话框

年份	月份	销售额	YEAR_	MONTH_	DATE_
1998	1	39.01	1998	1	JAN 1998
1998	2	44.24	1998	2	FEB 1998
1998	3	39.50	1998	3	MAR 1998
1998	4	38.25	1998	4	APR 1998
1998	5	38.04	1998	5	MAY 1998
1998	6	44.15	1998	6	JUN 1998
1998	7	43.52	1998	7	JUL 1998
1998	8	46.65	1998	8	AUG 1998
1998	9	42.30	1998	9	SEP 1998
1998	10	41.87	1998	10	OCT 1998

图 11-3 定义日期变量后的数据结果（部分）

11.1.3 创建时间序列

时间序列分析建立在序列平稳的条件上，判断序列是否平稳可以看它的均数方差是否不再随时间的变化而变化，自相关系数是否只与时间间隔有关而与所处时间无关。在时间序列分析中，为检验时间序列的平稳性，经常要用一阶差分、二阶差分，有时为选择一个合适的时间序列模型还要对原时间序列数据进行对数转换或平方转换等。这就需要在已经建立的时间序列数据文件中，再建立一个新的时间序列变量。

【例 11-3】 沿用例 11-1 的数据，为销售额创建一个时间序列。

第 1 步 创建时间序列的设置。

按“转换→创建时间序列”顺序打开“创建时间序列”对话框，将“销售额”变量移入右侧的“变量→新名称”框中，并在下面选择相应的函数。此处选择“中心移动平均值”，并将跨度设为 5，之后单击“变化量 (H)”按钮，设置情况如图 11-4 所示。



图 11-4 “创建时间序列”对话框

现对“函数”下拉列表中的方法解释如下。

- “差异”：产生差分序列。
- “季节性差异”：产生季节性差分序列，需要在“顺序”文本框中输入差分的阶。

- “中心移动平均值”：产生以当前值为中心的移动平均序列，需要在“跨度”文本框中输入宽度参数。
- “前移动平均值”：产生以当前值之前的相邻值计算的移动平均序列，需要在“跨度”文本框中输入宽度参数，在序列的开始处会产生和窗口宽度相等数目的缺失值。
- “运行中位数”：类似中心移动平均法，只不过此处计算的是相应的中位数。
- “累计求和”：计算累计和序列（当前值及所有历史值之和）。
- “延迟”：产生滞后序列。
- “提前”：产生领先序列。
- “平滑”：产生基于混合数据平滑法计算的平滑序列。

☆说明☆

- ◆ 通常讲的差分，是当前数据减去前一时间数据的含义，即差分的间隔为 1；而季节性差分，为当前“季节”减去前一“季节”的结果，差分的间隔和季节周期的选取有关，如果数据按天计，周期为周，则季节性差分间隔为 7。差分的阶是指差分的次数，1 阶差分为对原始数据做 1 次差分处理，2 阶差分为对 1 阶差分序列再做 1 次差分处理，等等。差分的阶和差分的间隔是两个不同的概念，差分序列必然会产生一定数量的缺失值，缺失值的数量 = 差分间隔 × 差分的阶。

第 2 步 结果及分析。

运行完成后，在数据文件中增加了 1 个按宽度为 5 的“中心移动平均值”计算出的时间序列变量，变量名为“销售额_1”。

第 3 步 作序列图分析。

按“分析→时间序列预测→序列图”顺序打开“序列图”对话框，并将“销售额”和“销售额_1”（第 2 步所生成的时间序列变量）移到右侧的“变量”框，并将已定义的日期变量设为“时间轴标签”，提交系统运行，其图形如图 11-5 所示，可看出销售额数据被平滑了。

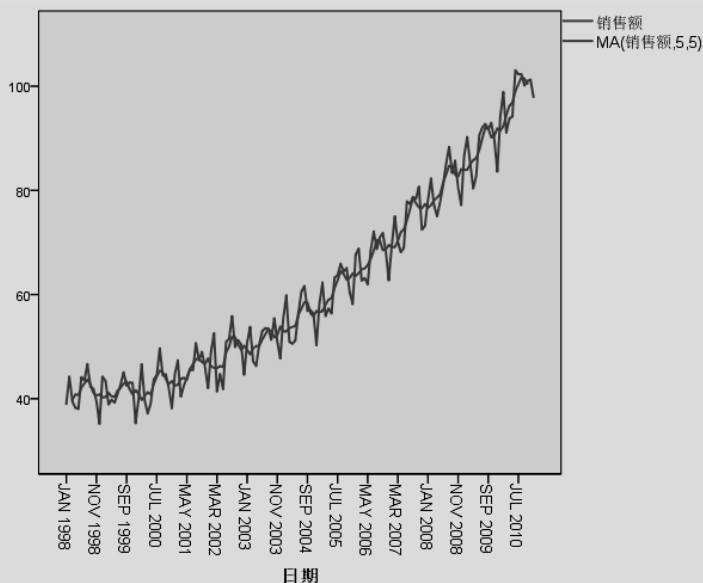


图 11-5 生成的序列图

11.2 指数平滑法

11.2.1 基本概念及统计原理

1. 基本概念

指数平滑法的思想来源于对移动平均预测法的改进。用当前值和历史值预测未来值时，移动平均法面临两个难题：其一是当前值和历史值同等权重不合理，一般而言，未来值总是和邻近时点的值关系更密切；其二是无法令人信服地确定窗口宽度，如使用 5 日移动平均数还是 15 日平均数难有定论，而且，如果使用 5 日移动平均数，那么 5 日之前的观察值等于赋予权重 0，而 5 日内的观察值均有相等权重 0.2，这也和实际情况相悖。指数平滑法的思想是以无穷大为宽度，各历史值的权重随时间的推移呈指数衰减，这样就解决了移动平均的两个难题。

2. 统计原理

指数平滑法用公式表达为

$$\hat{z}_{t+1} = \frac{\sum_{j=0}^{\infty} \theta^j z_{t-j}}{\sum_{j=0}^{\infty} \theta^j} = (1-\theta) \sum_{j=0}^{\infty} \theta^j z_{t-j} \tag{11.1}$$

式中， $0 \leq \theta \leq 1$ ； $j = 0, 1, 2, \cdots$ ； $t = 1, 2, \cdots$ ； $t > j$ 。

时间序列自身一般有随机波动、长期（线性或非线性）趋势和周期性（稳定性和不稳定性）波动三方面特征。

指数平滑法使用的模型很多，为了简洁明了，现将最常见的“简单”模型和“Holt 线性趋势”模型解释如下。

(1) 简单模型

简单模型是在移动平均法基础上发展而来的一次指数平滑法，它假定所研究的时间序列数据集无趋势或季节变化。其公式为

$$\hat{z}_{t+1} = \alpha z_t + (1-\alpha) \hat{z}_t \tag{11.2}$$

它改变了移动平均法用来预测的 N 个过去观测值中每一个权重都相等，而早于 $(t-N+1)$ 期的观测值的权数等于零，只适用于线性估计的局限。体现了对未来的估计，最近的观测值要比较早的观测值影响更大，在预测时应赋予更大权重的思想。如果用分量代替 \hat{z}_t ，则对式 (11.2) 展开可得到

$$\hat{z}_{t+1} = \alpha z_t + (1-\alpha)[\alpha z_{t-1} + (1-\alpha) \hat{z}_{t-1}] \tag{11.3}$$

进一步可将上式改写为

$$\hat{z}_{t+1} = \alpha z_t + \alpha(1-\alpha)z_{t-1} + \alpha(1-\alpha)^2 z_{t-2} + \cdots + \alpha(1-\alpha)^{N-1} \hat{z}_{t-N+1} \tag{11.4}$$

由式 (11.4) 可知，每一递推观测值的权数按指数规律递减，这就是指数平滑得名的原因。

(2) Holt 线性趋势模型

Holt 双参数线性指数平滑法适用于有线性趋势、无季节变化的时间序列的预测。它可以用不同的参数对原时间序列的趋势进行平滑，具有很大的灵活性。在此法中要用到两个参数 α 、 γ （在 0~1 之间取值）和三个方程式

$$\hat{z}_t = \alpha z_t + (1-\alpha)(\hat{z}_{t-1} + \hat{b}_{t-1}), \hat{z}_1 = z_1, 0 \leq \alpha \leq 1 \tag{11.5}$$

$$\hat{b}_t = \gamma(\hat{z}_t - \hat{z}_{t-1}) + (1-\gamma)\hat{b}_{t-1}, \hat{b}_1 = 0, 0 \leq \gamma \leq 1 \tag{11.6}$$

$$\hat{z}_{t+m} = \hat{z}_t + \hat{b}_t m \tag{11.7}$$

式 (11.5) 利用前一期的趋势值 \hat{b}_{t-1} 直接修正 \hat{z}_t ，即将 \hat{b}_{t-1} 加在前一平滑值 \hat{z}_{t-1} 上，用来消除滞后，且使 \hat{z}_t 值近似达到最新数据值 z_t 。式 (11.6) 用来修正趋势值 \hat{b}_t ，趋势值用相邻两次平滑值之差来表示。式 (11.7) 进行预测，预测值为基础值加上趋势值乘以预测超前期数。

11.2.2 指数平滑法 SPSS 实例分析

【例 11-4】表 11.2 是我国 1996~2015 年私人汽车拥有量数据，试用指数平滑法对全国私人汽车拥有量进行预测分析。（数据来源：《中国统计年鉴》，2016 年；参见数据文件：data11-2.sav。）

表 11.2 全国 1996~2015 年全国私人汽车拥有量（单位：万辆）

年份	私人汽车拥有量	年份	私人汽车拥有量
1996	289.67	2006	2333.32
1997	358.36	2007	2876.22
1998	423.65	2008	3501.39
1999	533.88	2009	4574.91
2000	625.33	2010	5938.71
2001	770.78	2011	7326.79
2002	968.98	2012	8838.6
2003	1219.23	2013	10501.68
2004	1481.66	2014	12339.36
2005	1848.07	2015	14099.1

第 1 步 数据组织。

将数据组织成 2 列，一列是“年份”，另一列是“私人汽车拥有量”，输入数据并保存。

第 2 步 分析。

看用指数平滑法处理是否恰当。按 11.1.3 节所述创建私人汽车拥有量的序列图，如图 11-6 所示。从此图可以看出，私人汽车拥有量呈逐年增加趋势，开始增长较慢，然后变快，近似线性趋势，也可以说呈增长的线性趋势，或者用指数趋势描述更准确。所以可选用指数平滑法进行处理。

第 3 步 定义日期变量。

按 11.1.2 节所示将“年份”定义为日期变量。

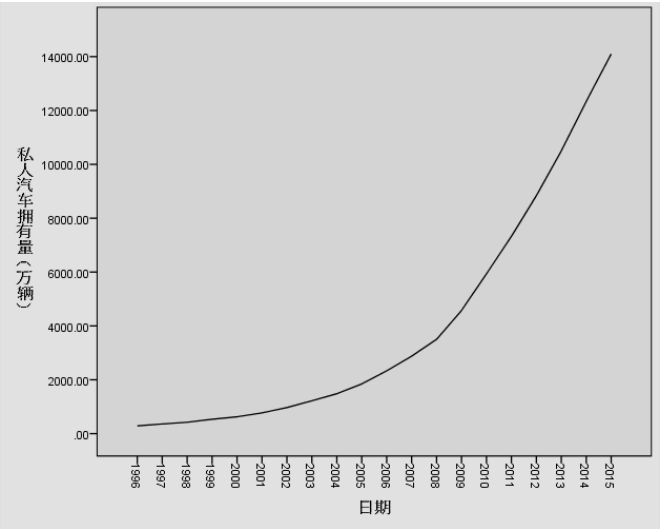


图 11-6 私人汽车拥有量的序列图

第 4 步 指数平滑法设置。

(1) 按“分析→时间序列预测→创建传统模型”顺序打开“时间序列建模器”对话框，并按图 11-7 所示进行设置，现将对话框中的各项解释如下。



图 11-7 “时间序列建模器”对话框

(2) “变量”选项卡设置：其中包括要选择的因变量，本例中将“私人汽车拥有量”设为自变量。

① 时间序列“方法”的选择：其中包括 3 种方法，此处设置为“指数平滑”法，对几种方法的解释如下。

- 专家建模器：专家建模器会自动查找每个相依序列的最佳拟合模型，专家建模器既考虑指数平滑法模型，也考虑 ARIMA 模型。
 - 指数平滑法：使用此选项可指定定制的指数平滑法模型。可以从各种指数平滑法模型中进行选择，它们在处理趋势和季节性上有所不同。
 - ARIMA：用此选项可指定定制的 ARIMA 模型。其中包含显式指定自回归的阶和移动平均的阶，以及差分度。可以包含自变量（预测变量）并为它们当中的任何一个或全部定义转换函数。
- ② “条件”设置：单击“条件(C)…”按钮，打开“时间序列建模器：指数平滑条件”对话框，这里选择“霍尔特线性趋势”模型，设置如图 11-8 所示。



图 11-8 “时间序列建模器：指数平滑条件”对话框

图 11-8 中包含了两个大的选项组“模型类型”和“因变量转换”，现对各项解释如下。

指数平滑法模型（Gardner, 1985）分为季节性模型和非季节性模型。季节性模型只有在为活动数据集定义了周期时才可用。

- “简单”模型：该模型适用于没有趋势或季节性的序列。
- “霍尔特线性趋势”模型：该模型适用于具有线性趋势且没有季节性的序列。其平滑参数是水平和趋势，不受相互之间值的约束。霍尔特（Holt）模型比布朗（Brown）模型更通用，但在计算大序列时花的时间更长。
- “布朗线性趋势”模型：该模型适用于具有线性趋势且没有季节性的序列。其平滑参数是水平和趋势，并假定二者等同。因此，Brown 模型是 Holt 模型的特例。
- “衰减趋势”模型：此模型适用于具有线性趋势的序列，该线性趋势正逐渐消失且没有季节性。
- “简单季节性”模型：该模型适用于没有趋势且季节性影响随时间变动保持恒定的序列，其平滑参数是水平和季节。
- “温特斯加性”模型：该模型适用于具有线性趋势和不依赖于序列水平的季节性效应的序列。其平滑参数是水平、趋势和季节。
- “温特斯乘性”模型：该模型适用于具有线性趋势和依赖于序列水平的季节性效应的序列。

- 当前周期长度：指示当前为活动数据集定义的周期长度。
 - 因变量转换：可以指定在建模之前对每个因变量执行的转换，包括不执行转换、平方根转换、自然对数转换几种方法。
- (3) “统计”选项卡设置：设置如图 11-9 所示。

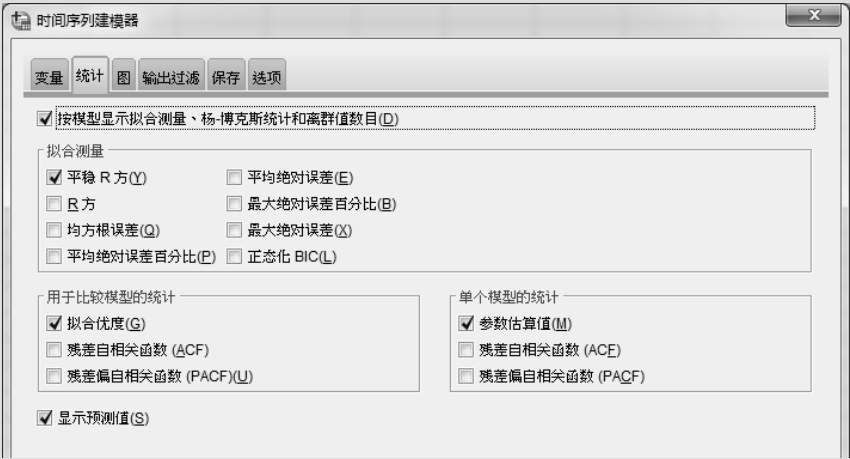


图 11-9 “时间序列建模器：统计”选项卡设置

现对其中各项解释如下。

- ① “按模型显示拟合测量、杨-博克斯统计和离群值数目”选项：选中此选项可显示包含每个估计模型的所选拟合测量、杨-博克斯（Ljung-Box）值以及离群值数目的表。
- ② “拟合测量”选项组用于选择拟合度量的方法，主要包括以下几种方法。
- “平稳 R 方”度量：将模型的平稳部分与简单均值模型相比较的测量。当具有趋势或季节性模式时，该度量适用于普通 R 方。固定的 R 方可以是 $-\infty \sim 1$ 范围中的负值。负值表示考虑中的模型比基线模型差，正值表示考虑中的模型比基线模型好。
 - “R 方”度量：总变动在由模型解释的序列中的比例估计。当序列很平稳时，此度量最有用。
 - “均方根误差”度量：度量因变量序列与其模型预测水平的相差程度，用和因变量序列相同的单位表示。
 - “平均绝对误差百分比”度量：度量因变量序列与其模型预测水平的相差程度。它与使用的单位无关，因此可用于比较具有不同单位的序列。
 - “平均绝对误差”度量：度量序列与其模型预测水平的差别程度。
 - “最大绝对误差百分比”度量：最大的预测误差，以百分比表示。
 - “最大绝对误差”度量：最大的预测误差，以和因变量序列相同的单位表示。
 - “正态化 BIC”度量：尝试代表模型复杂性的模型整体拟合的一般度量。它是基于均方误差的分数，包括模型中参数数量的罚分和序列长度。罚分去除了具有更多参数的模型优势，从而可以容易地比较相同序列的不同模型的统计量。
- ③ “用于比较模型的统计”选项组：用于控制输出的统计信息表，每个选项分别生成单独的表，可以选择以下选项中的一个或多个。
- “拟合优度”选项：显示包含固定的 R 方等拟合优度表。
 - “残差自相关函数”选项：所有估计模型中残差的自相关摘要统计和百分位表。

- “残差偏自相关函数”选项：所有估计模型中残差的部分自相关摘要统计和百分位表。
- ④ “单个模型的统计”选项组：用于控制如何显示包含每个估计模型的详细信息的表。每个选项分别生成单独的表。可以选择以下选项中的一个或多个。
 - “参数估算值”选项：显示每个估计模型的参数估计值的表。为指数平滑法和 ARIMA 模型显示不同的表。如果存在离群值，则它们的参数估计值也将在单独的表中显示。
 - “残差自相关函数”选项：按每个估计模型的延迟显示残差自相关表。该表包含自相关的置信区间。
 - “残差偏自相关函数”选项：按每个估计模型的延迟显示残差部分自相关表。该表包含部分自相关的置信区间。
- ⑤ “显示预测值”选项：显示每个估计模型的模型预测和置信区间的表。预测期在“选项”选项卡中设置。
- (4) “图”选项卡的设置：在“图”选项卡中选择“序列”、“实测值”、“预测值”和“拟合值”四项，其中各项的解释与“统计”选项卡类似。
- (5) “保存”选项卡的设置：如图 11-10 所示，将“预测值”保存到数据文件中，预测期在“选项”选项卡中设置。可以保存的变量有“预测值”、“置信区间”的上限和下限、“噪声残值”4 项。



图 11-10 “时间序列建模器：保存”选项卡

- (6) “选项”选项卡设置：具体设置如图 11-11 所示。该选项卡用于设置预测期、指定缺失值的处理方法、设置置信区间宽度、指定模型标识的定制前缀 ACF 和 PACF 输出中的显示标签最大数。此例中我们设置预测期到 2017 年，其他为默认设置。



图 11-11 “时间序列建模器：选项”选项卡

在“预测期”选项组中有以下两个选项。

- “评估期结束后的第一个个案到活动数据集中的最后一个个案”选项：如果估计期在活动数据集中的最后一个个案之前结束，而您需要直到最后一个个案的预测值，则请选择此选项。此选项通常用来生成保持期的预测，以便将模型预测与实际值子集进行比较。
- “评估期结束后的第一个个案到指定日期之间的个案”选项：选择此选项可显式指定预测期的结束。此选项通常用于在实际序列结束后生成预测，在“日期”网格中为所有单元格输入值。

所有这些选项卡设置完成后提交系统运行。

第 5 步 主要结果及分析。

主要结果如表 11.3 ~ 11.7 及图 11-12、图 11-13 所示，具体分析如下。

(1) 表 11.3 是模型的描述表，表示对“私人汽车拥有量”变量进行指数平滑法处理，使用的是“霍尔特”模型。

表 11.3 模型描述表

			模型类型
模型 ID	私人汽车拥有量 (万辆)	模型_1	霍尔特

(2) 表 11.4 是模型的拟合情况表，包含了 8 个拟合情况度量指标，其中“平稳 R 方”值为 -0.642，“R 方”值为 0.999，并给出了每个度量模型的百分位数。

(3) 表 11.5 是模型统计量表，从中可以看出模型的“平稳 R 方”值为 -0.642，另外还给出

了拟合统计量及杨-博克斯统计情况,可看出其显著性为 0.329。此外,所有数据中没有离群值(孤立点)。

表 11.4 模型拟合情况

拟合统计	平均值	标准误差	最小值	最大值	百分位数						
					5	10	25	50	75	90	95
平稳R方	-.642	.	-.642	-.642	-.642	-.642	-.642	-.642	-.642	-.642	-.642
R方	.999	.	.999	.999	.999	.999	.999	.999	.999	.999	.999
RMSE	150.346	.	150.346	150.346	150.346	150.346	150.346	150.346	150.346	150.346	150.346
MAPE	3.234	.	3.234	3.234	3.234	3.234	3.234	3.234	3.234	3.234	3.234
MaxAPE	9.801	.	9.801	9.801	9.801	9.801	9.801	9.801	9.801	9.801	9.801
MAE	94.594	.	94.594	94.594	94.594	94.594	94.594	94.594	94.594	94.594	94.594
MaxAE	448.395	.	448.395	448.395	448.395	448.395	448.395	448.395	448.395	448.395	448.395
正态化 BIC	10.325	.	10.325	10.325	10.325	10.325	10.325	10.325	10.325	10.325	10.325

表 11.5 模型统计量

模型	预测变量数	模型拟合度统计	杨-博克斯 Q(18)			离群值数
		平稳R方	统计	DF	显著性	
私人汽车拥有量(万辆)-模型_1	0	-.642	17.908	16	.329	0

(4) 表 11.6 是指数平滑法拟合的模型参数表,可以看出 α 取值为 1.000, γ 取值为 1.000, 从对应的显著性概率值可看出均较小,说明两参数具有一定的显著意义。则根据式 (11.5) 可得 $\hat{z}_t = \alpha z_t$ 。

表 11.6 指数平滑法模型参数

模型			估算	标准误差	t	显著性
私人汽车拥有量(万辆)-模型_1	不转换	Alpha (水平)	1.000	.325	3.077	.006
		Gamma (趋势)	1.000	.505	1.980	.063

(5) 表 11.7 是预测情况表,表中给出了 2016~2017 年“私人汽车拥有量”变量的预测值、上区间和下区间值。

表 11.7 预测数据

模型		2016	2017
私人汽车拥有量(万辆)-模型_1	预测	15858.98	17618.82
	UCL	16174.84	18324.88
	LCL	15543.11	16912.76

对于每个模型,预测从所请求估算期范围内的最后一个非缺失值之后开始,并结束于最后一个所有预测变量都有可用的非缺失值的周期,或者在所请求预测期的结束日期结束,以较早者为准。

(6) 图 11-12 是观测值与预测值的序列图。

图 11-12 是实测值、拟合值和预测值的序列图,可发现该模型对历史数据的拟合较好。

(7) 图 11-13 是按指数平滑法预测的 2016~2017 年“私人汽车拥有量”保存在文件中的数据。

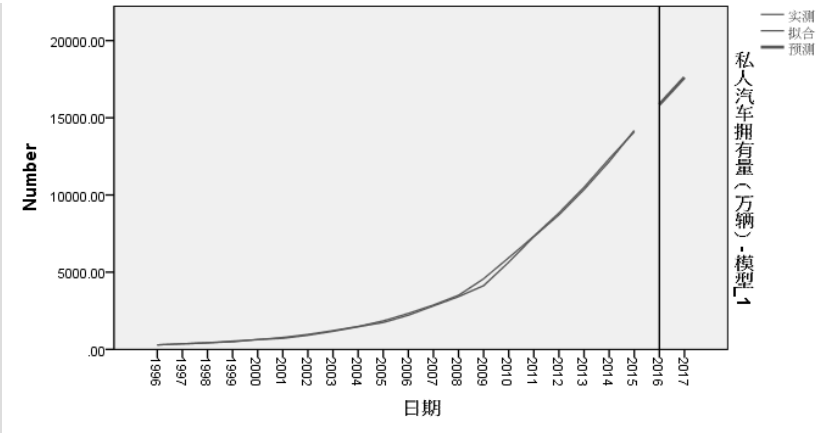


图 11-12 实测值、拟合值与预测值情况

年份	私人汽车拥有量（万辆）	YEAR_	DATE_	预测_私人汽车拥有量（万辆）_模型_1
2002	968.98	2002	2002	916.18
2003	1219.23	2003	2003	1167.16
2004	1481.66	2004	2004	1469.46
2005	1848.07	2005	2005	1744.10
2006	2333.32	2006	2006	2214.40
2007	2876.22	2007	2007	2818.51
2008	3501.39	2008	2008	3419.12
2009	4574.91	2009	2009	4126.52
2010	5938.71	2010	2010	5648.09
2011	7326.79	2011	2011	7302.45
2012	8838.60	2012	2012	8714.97
2013	10501.68	2013	2013	10350.32
2014	12339.36	2014	2014	12164.69
2015	14099.10	2015	2015	14176.96
-	-	2016	2016	15858.98
-	-	2017	2017	17618.82

图 11-13 数据文件中的保存结果

11.3 ARIMA 模型

11.3.1 基本概念及统计原理

1. 基本概念

在预测中，对于平稳的时间序列，可用自回归移动平均（AutoRegressive Moving Average, ARMA）模型及特殊情况的自回归（AutoRegressive, AR）模型、移动平均（Moving Average, MA）模型等来拟合，预测该时间序列的未来值，但在实际的经济预测中，随机数据序列往往都是非平稳的，此时就需要对该随机数据序列进行差分运算，进而得到 ARMA 模型的推广——ARIMA 模型。

ARIMA 模型全称综合自回归移动平均（AutoRegressive Integrated Moving Average）模型，简记为 ARIMA(p, d, q) 模型，是由 Box 和 Jenkins 于 20 世纪 70 年代初提出的著名时间序列预测模型，

又称为 Box-Jenkins 模型。其中 AR 是自回归, p 为自回归阶数; MA 为移动平均, q 为移动平均阶数; d 为时间序列成为平稳时间序列时所做的差分次数。ARIMA(p, d, q)模型的实质就是差分运算与 ARMA(p, q)模型的组合, 即 ARMA(p, q)模型经 d 次差分后, 便为 ARIMA(p, d, q)。

2. 统计原理

(1) ARMA 过程

设 $\{x_t\}$ 为零均值平稳序列, $\{a_t\}$ 为白噪声, $Ex_t a_s = 0 (t < s)$ (Ex_t 称为时间序列的均值序列, 它是和时间有关的序列), 满足

$$x_t - \varphi_1 x_{t-1} - \varphi_2 x_{t-2} - \cdots - \varphi_p x_{t-p} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \quad (11.8)$$

式中, x_{t-p} 表示比时刻 t 滞后 p 个时间的数值。则 $\{x_t\}$ 为 p 阶自回归— q 阶滑动平均过程, 简记为 ARMA(p, q)。 $\{x_t\}$ 称为 ARMA(p, q)序列, 非负整数 p, q 分别称为自回归阶数和滑动平均阶数, 参数 $\varphi_1, \varphi_2, \cdots, \varphi_p$ 称为自回归系数, $\theta_1, \theta_2, \cdots, \theta_q$ 称为滑动平均系数。

当 $p = 0$ 时, 则为 ARMA(0, q)模型

$$x_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \quad (11.9)$$

称为 q 阶滑动平均模型, 记为 MA(q)。当 $q = 0$ 时, 则为 ARMA($p, 0$)模型

$$x_t - \varphi_1 x_{t-1} - \varphi_2 x_{t-2} - \cdots - \varphi_p x_{t-p} = a_t \quad (11.10)$$

称为 p 阶自回归模型, 记为 AR(p)。

引入后移(延迟)算子 B , 令 $B^k x_t = x_{t-k}$, $B^k a_t = a_{t-k}$, $B^k c = c$ (c 为常数), 并令

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p, \quad \theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \quad (11.11)$$

则 ARMA(p, q)模型简记为

$$\varphi(B)x_t = \theta(B)a_t \quad \text{或} \quad x_t = \varphi^{-1}(B)\theta(B)a_t \quad (11.12)$$

(2) ARMA 模型的识别

设 ACF 代表 $\{x_t\}$ 的自相关函数, PACF 代表 $\{x_t\}$ 的偏自相关函数。根据 Box-Jenkins 提出的方法, 用样本的自相关函数 (ACF) 和偏自相关函数 (PACF) 的截尾性来初步识别 ARMA 模型的阶数。具体如表 11.8 所示。

表 11.8 ARMA 模型的识别

模型	自相关函数 (ACF)	偏自相关函数 (PACF)
AR(p)	拖尾	p 阶截尾
MA(q)	q 阶截尾	拖尾
ARMA(p, q)	拖尾	拖尾

☆说明☆

- ◆ 所谓拖尾是自相关系数或偏相关系数逐步趋向于 0, 这个趋向过程有不同的表现形式, 有几何型的衰减, 有正弦波式的衰减; 而所谓截尾是指从某阶后自相关或偏相关系数为 0。

表 11.8 是我们判断时间序列模型的形式和进行模型拟合的重要依据。当为 ARMA(p, q)序列时, 首先可以根据经验给出 p, q 的初步识别, 然后通过模型诊断反复识别, 找出最优的 p, q 组合来确定。

(3) 非平稳时间序列——ARIMA 过程

定义一阶差分算子 ∇ 为 $\nabla z_t = z_t - z_{t-1}$ ，则差分算子 ∇ 和延迟算子 B 有关系式

$$\nabla = 1 - B, \quad \nabla^2 = (1 - B)^2, \quad \nabla^d = (1 - B)^d \tag{11.13}$$

式中， d 为差分的阶。

季节差分：季节差分中 k 一般取一个周期，如对于月度数据 $k = 12$ ，对于季度数据 $k = 4$ 等，季节差分算子为

$$\nabla_k = x_t - x_{t-k} \tag{11.14}$$

设 $\{z_t\}$ 为非平稳序列， $\{x_t\}$ 为 ARMA(p, q) 序列，存在正整数 d ，使得 $x_t = \nabla^d z_t, t > d$ ，则有

$$\varphi(B)(1 - B)^d z_t = \theta(B)a_t \tag{11.15}$$

称此模型为综合自回归移动平均模型，记为 ARIMA(p, d, q)。

(4) 季节 ARIMA 模型

时间序列常呈周期性变化，或称为季节性趋势。用变通的 ARIMA 模型处理这种季节性趋势会导致参数过多，模型复杂。季节性乘积模型可以得到参数简约的模型。季节性乘积模型表示为 ARIMA($p, d, q, \text{sp}, \text{sd}, \text{sq}$) (或 ARIMA(p, d, q) $\times(\text{sp}, \text{sd}, \text{sq})_k$)。其中 sp 表示季节模型的自回归系数；sd 表示季节差分的阶数，通常为一阶季节差分；sq 表示季节模型的移动平均参数。如是月度资料，要描述年度特征，则 sd=12；如是日志资料，欲描述每周特征，则 sd=7。

3. ARIMA 建模步骤

ARIMA 建模实际上包括 3 个阶段，即模型识别阶段、参数估计和检验阶段、预测应用阶段。其中前两个阶段可能需要反复进行。

ARIMA 模型的识别就是判断 $p, d, q, \text{sp}, \text{sd}, \text{sq}$ 的阶，主要依靠自相关函数 (ACF) 和偏自相关函数 (PACF) 图来初步判断和估计。一个识别良好的模型应该有两个要素：一是模型的残差为白噪声序列，需要通过残差白噪声检验，二是模型参数的简约性和拟合优度指标的优良性 (如 对数似然值较大，AIC 和 BIC 较小) 方面取得平衡，还有一点需要注意，模型的形式应该易于理解。

11.3.2 ARIMA 实例分析

【例 11-5】 表 11.9 是某加油站 55 天的燃油剩余数据，其中正值表示燃油有剩余，负值表示燃油不足，要求对此序列拟合时间序列模型并进行分析。(参见数据文件：data11-3.sav。)

第 1 步 数据组织。

将数据组织成两列，一列是“天数”，另一列是“燃油量”，输入数据并保存，并以“天数”定义日期变量，会新增一个名为“DATE_”的变量。

第 2 步 观察数据序列的性质。

(1) 作序列图，观察数据序列的特点。按“分析→时间序列预测→序列图”的顺序打开“序列图”对话框，将“油料量”设置为变量，并将所生成的日期新变量“DATE_”设为时间标签轴，生成如图 11-14 所示的序列图。可以看出数据序列在 0 上下振荡，且无规律，可能是平稳的时间序列。

(2) 再做自相关图和偏自相关图进一步分析。按“分析→时间序列预测→自相关”顺序打开“自相关”对话框，将“燃油量”选入“变量”框，并在“输出”选项组中将“自相关”和

“偏自相关”同时选上，输出结果如图 11-15 和图 11-16 所示，图形横坐标“滞后编号”表示延迟数。从图 11-15 可以看出，自相关函数呈现出比较典型的拖尾性，说明数据自相关性随时间间隔下降。从图 11-16 可以看出，除了延迟 1 阶的偏自相关系数在 2 倍标准差范围之外，其他除数的偏自相关系数都在 2 倍标准差范围内波动。根据这个特点可以判断该序列具有短期相关性，进一步确定序列平稳。同时，可以认为该序列偏自相关函数 1 阶截尾。

表 11.9 某加油站燃油剩余数据（单位：t）

天	燃油数据	天	燃油数据	天	燃油数据	天	燃油数据
1	92	15	78	29	-60	43	15
2	-85	16	-98	30	-50	44	20
3	80	17	-9	31	30	45	15
4	12	18	75	32	-10	46	90
5	10	19	65	33	3	47	15
6	3	20	80	34	-65	48	-10
7	-1	21	-20	35	10	49	-8
8	-2	22	-85	36	8	50	8
9	0	23	0	37	-10	51	0
10	-90	24	1	38	10	52	25
11	100	25	150	39	-25	53	-120
12	-40	26	-100	40	90	54	70
13	-2	27	135	41	-30	55	-10
14	20	28	-70	42	-32		

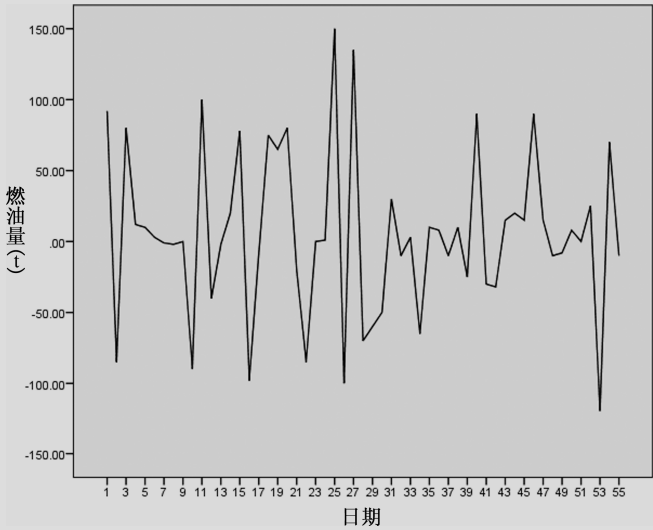


图 11-14 所生成的时间序列图

综合该序列自相关函数和偏自相关函数的性质，根据表 11.8 的模型识别规则，可以拟合模型为 AR(1)，即 ARIMA(1, 0, 0)。

第 3 步 模型拟合。

（1）按“分析→时间序列预测→创建模型”顺序打开“时间序列建模器”对话框，将“燃油量”选入“因变量”框。设置过程与图 11-17 类似，并选择“方法”下的“ARIMA”模型。

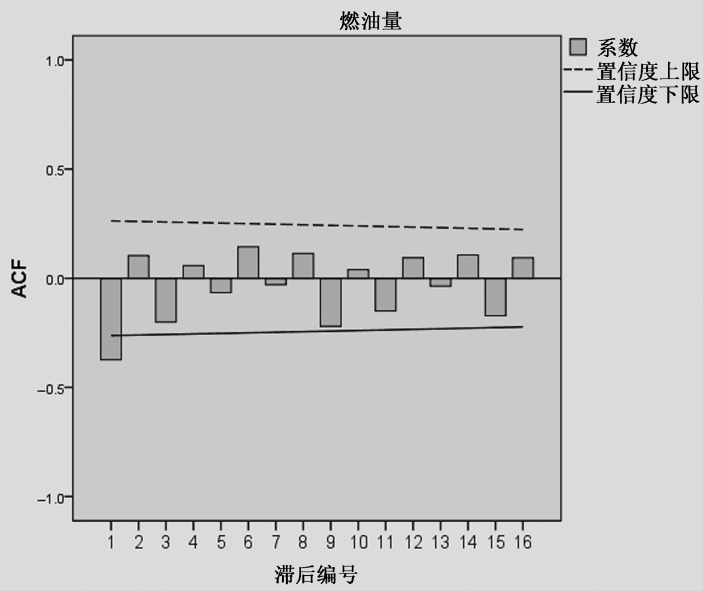


图 11-15 燃油量的自相关图

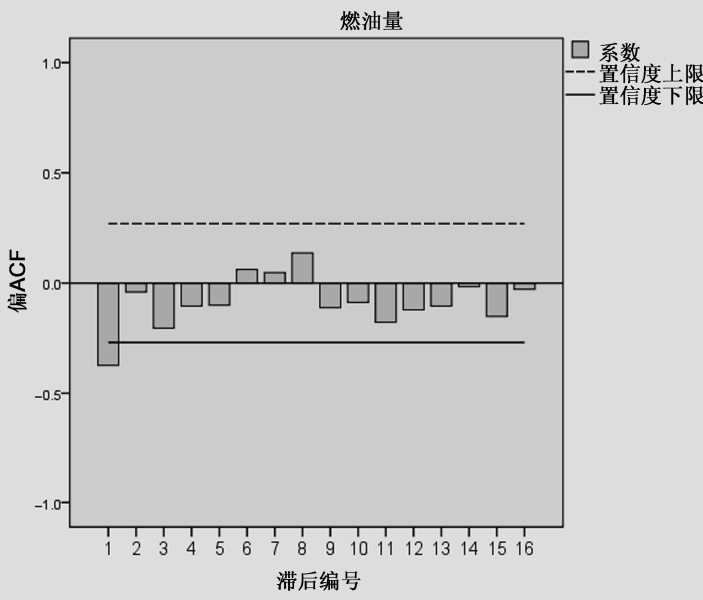


图 11-16 燃油量的偏自相关图

(2) “条件”对话框设置。单击“方法”右边的“条件 (C) ...”按钮，打开“时间序列建模器：ARIMA 条件”对话框，并按如图 11-17 所示进行设置。在“ARIMA 阶”框中需设置“非季节性”参数：自回归的阶 p 、差值的阶 d 和移动平均值 q 。如果时间序列有季节性因素，还需设置“季节性”参数 sp 、 sd 和 sq 。由于经过前面的分析，此例是 ARIMA(1, 0, 0) 模型，且无季节性影响，则只需将自回归的阶数设为 1，其余均为 0。



图 11-17 “时间序列建模器：ARIMA 条件”对话框

(3) “统计”选项卡的设置：“统计”选项卡如图 11-9 所示，将“按模型显示拟合测量、杨-博克斯统计和离群值数目”、“R 方”、“正态化 BIC”、“拟合优度”、“参数估算值”选上。

(4) “图”选项卡的设置：在其中将“序列”、“残差自相关函数”、“残差偏自相关函数”、“实测值”和“预测值”这些选项选上。

其他选项卡的设置读者可参照例 11-4 进行。

第 4 步 主要结果及分析。

主要结果如表 11.10、表 11.11 及图 11-18 所示，具体分析如下。

(1) 表 11.10 是模型的统计表，列出了模型拟合的一些统计量，包括决定系数（R 方）、正态化 BIC 值、杨-博克斯统计量值，从结果看，拟合效果不太理想，决定系数的值偏小，而且从显著性概率值大于 0.05 来看，杨-博克斯统计量的观测值也不显著。

表 11.10 模型统计

模型	预测变量数	模型拟合度统计		杨-博克斯 Q(18)			离群值数
		R 方	正态化 BIC	统计	DF	显著性	
燃油量-模型_1	0	.139	8.170	14.688	17	.618	0

(2) 表 11.11 是 ARIMA 模型参数表，从结果可以看出，AR(1)模型的参数为-0.382，参数是显著的，常数项为 4.69，不显著，这里仍然保留常数项。从结果来看，其拟合模型为 $x_t - 0.382x_{t-1} = 4.69 + a_t$ 。

表 11.11 ARIMA 模型参数表

				估算	标准误差	t	显著性
燃油量-模型_1	燃油量	不转换	常量	4.690	5.399	.869	.389
			AR 延迟 1	-.382	.127	-3.020	.004

(3) 图 11-18 是 ARIMA(1, 0, 0)模型拟合残差的自相关函数和偏自相关函数图, 可以看出, 残差的自相关和偏自相关函数都是 0 阶截尾的, 因而残差是一个不含相关性的白噪声序列。因此, 序列的相关性都已经充分拟合了。

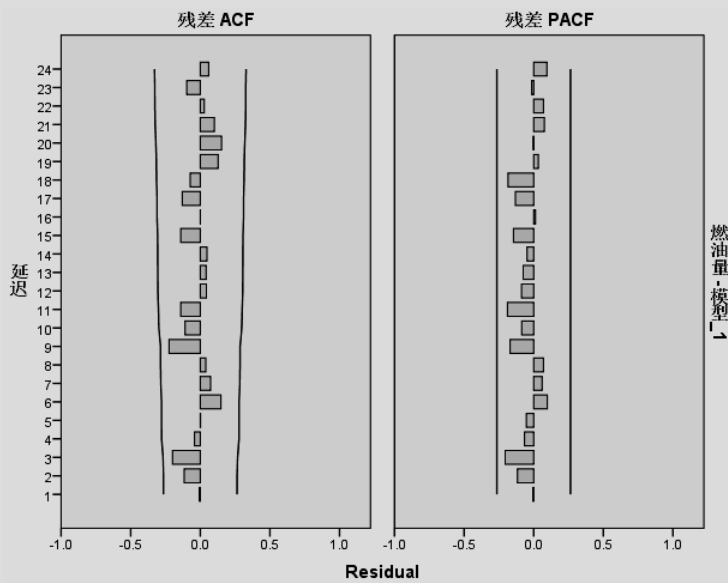


图 11-18 ARIMA(1, 0, 0)拟合残差的自相关和偏自相关函数图

☆说明☆

- (1) 从例 11-5 所拟合的结果来看是不够理想的, 读者可以通过调整自回归的阶和滑动平均的阶来选择更好的拟合模型。
- (2) 当然在求 ARIMA 拟合模型时, 还可以进行预测并将预测值保留下来, 具体方法请参看例 11-4。

【例 11-6】表 11.12 是我国 1992 年 1 月至 2002 年 12 月出口激光唱片的量 (单位: 万张), 试对该序列数据进行 ARIMA 模型拟合。(参见数据文件: data11-4.sav.)

表 11.12 1992 年 1 月~2002 年 12 月出口激光唱片数据

	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
1992	6.14	4.44	7.79	7.76	11.19	14.55	10.18	11.24	21.94	31.31	26.51	27.31
1993	13.55	23.57	19.43	11.19	10.27	19.83	16.07	16.93	30.78	29.74	54.12	60.35
1994	23.29	34.04	44.96	34.79	23.31	34.29	49.82	35.45	59.61	75.65	98.12	106.99
1995	65.31	40.26	71.92	64.4	67.7	59.71	56.27	93.45	100.15	152.06	109.41	104.12
1996	61.89	109.04	56.91	62.89	56.61	118.28	95.66	96.71	141.2	166.87	108.74	131.12
1997	66.08	46.37	72.36	79.73	100.52	130.86	148.91	134.42	149.72	253.28	211.59	209.38
1998	115.46	106.18	145.03	161.18	201.14	233.9	236.62	242.76	279.02	310.78	297.47	239.84
1999	90.01	106.27	204.16	157.23	213.89	225.72	198.27	234.59	307.01	387.37	335.57	374.88
2000	251.63	159.84	280.44	294.89	299.12	354.4	375.79	432.15	482	488.42	479.92	419.28
2001	247.59	225.39	294.44	332.77	331	399.68	418.95	485.22	544.71	618.88	503.27	334.58
2002	302.78	290.44	348.39	473.8	499.43	544.73	600.39	646.22	721.91	727.57	589.53	383.28

第 1 步 数据组织。

将数据组织成 3 列，第 1 列是“年份”，第 2 列是“月份”，第 3 列是“出口量”，并按“年份、月份”形式定义日期变量。

第 2 步 观察数据序列的性质。

先作序列图，观察数据序列的特点，如图 11-19 所示。可以看出，出口量序列图有明显的上升趋势，而且一年内 1 月出口量小，6 月出口量大，周而复始，因此该序列有趋势和季节规律，是不平稳的时间序列数据。（通过自相关和偏自相关函数图也可看出均没有衰减到 0，是不平稳的时间序列数据。）

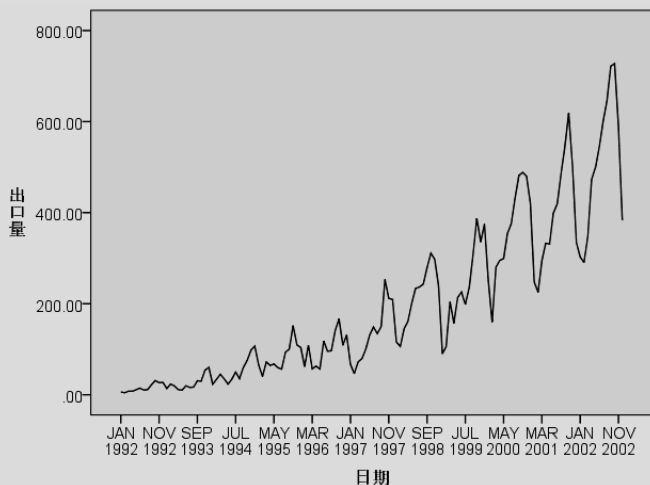


图 11-19 激光唱片出口量的序列图

由于数据中既有趋势信息，也有季节信息，因此需要对数据做一阶差异和一阶季节差异处理，具体方法是在“序列图”对话框的“转换”选项组中选择“自然对数转换”、“异差”和“季节性差异”，并在后面的文本框中制定差分的阶数为 1 阶，其序列图如图 11-20 所示，可以看出序列基本平稳。

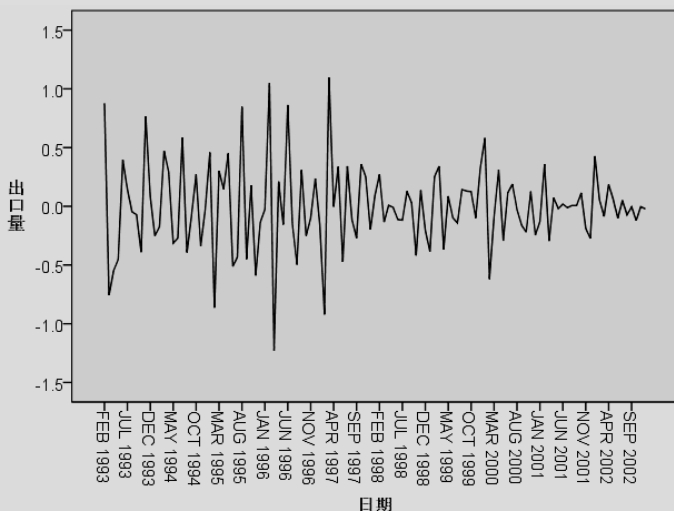


图 11-20 经过对数转换、差异和季节性差异后的序列图

下面需要用自相关函数和偏自相关函数验证序列的平稳性，此处仍需在“序列图”对话框的“转换”选项组中选择“自然对数转换”、“异差”和“季节性差异”，进行 1 阶差异转换，结果如图 11-21 和图 11-22 所示。可以看出，序列基本平稳，只是数据对于季节还有效应，因此需考虑季节自回归参数 sp 和季节移动平均系数 sq 。

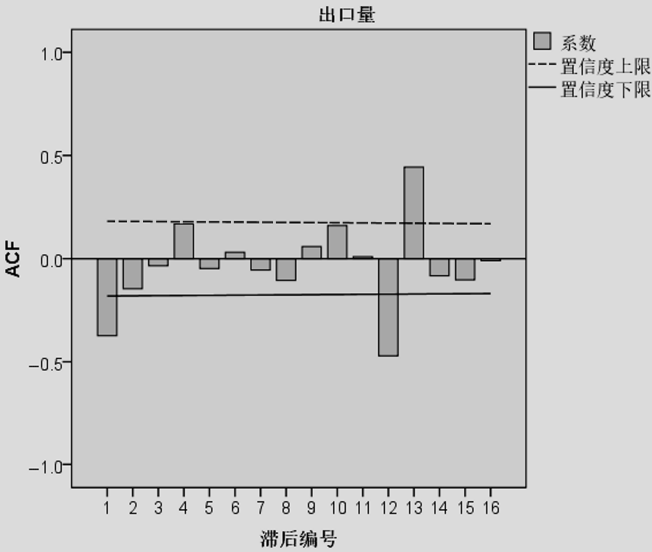


图 11-21 自相关函数图

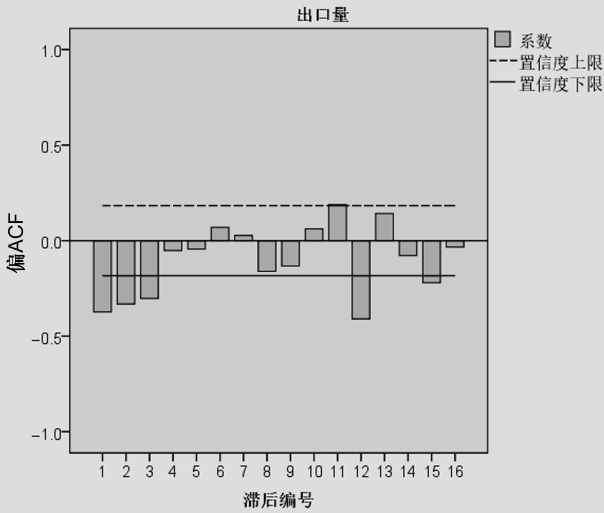


图 11-22 偏自相关函数图

第 3 步 ARIMA 模型的定阶。

接下来需要对 ARIMA 模型定阶，即确定自回归系数、移动平均系数、差异阶数、季节自回归系数、季节差异阶数、季节移动平均系数。从图 11-21 和图 11-22 可以基本判断自相关函数为 1 阶截尾，并且有 1 阶季节自相关，而偏自相关函数为拖尾，有 1 阶季节偏相关。此时可以估计，

模型可能是 ARIMA(1, 1, 0, 0, 1, 0), ARIMA(1, 1, 0, 0, 1, 1), ARIMA(1, 1, 0, 1, 1, 0), ARIMA(1, 1, 0, 1, 1, 1), ARIMA(1, 1, 1, 1, 1, 0)等模型中的一个, 具体选择哪一个模型, 则需要分别对模型进行拟合, 根据拟合效果的情况, 即决定系数和正态化 BIC 的值进行判断, 决定系数越高, BIC 值越小的模型拟合效果就越好, 最终需要在若干个模型中选择一个最优的模型, 这就是模型的定阶。

下面就来看拟合结果, 为了显示简洁, 特制作了各个模型的拟合效果表, 如表 11.13 所示。根据比较结果, 显然 ARIMA(1, 1, 1, 1, 1, 0)拟合效果最好, 下面就以此模型进行拟合, 看拟合的效果。

表 11.13 各种模型的拟合效果表

模型	决定系数 (R 方)	正态化 BIC
ARIMA(1, 1, 0, 0, 1, 0)	0.944	7.538
ARIMA(1, 1, 0, 0, 1, 1)	0.949	7.489
ARIMA(1, 1, 0, 1, 1, 0)	0.949	7.49
ARIMA(1, 1, 0, 1, 1, 1)	0.949	7.539
ARIMA(1, 1, 1, 0, 1, 1)	0.957	7.375
ARIMA(1, 1, 1, 1, 1, 0)	0.958	7.353
ARIMA(1, 1, 1, 1, 1, 1)	0.933	7.862

第 4 步 最优模型拟合主要结果及分析。

主要结果如表 11.14、表 11.15 及图 11-23、图 11-24 所示, 具体分析如下。

(1) 表 11.14 是模型的统计表, 从中可以看出模型的拟合效果比较理想, 决定系数 (R 方) 达到了 0.958, 说明模型可解释原序列 95.8%的信息, 正态化 BIC 值也比较小, 杨-博克斯 (Ljung-Box) 统计量的值也是显著的, 这些都说明用 ARIMA(1, 1, 1, 1, 1, 0)模型能很好地拟合时间序列数据。

表 11.14 模型统计

模型	预测变量数	模型拟合度统计		杨-博克斯 Q(18)			离群值数
		R 方	正态化 BIC	统计	DF	显著性	
出口量-模型_1	0	.958	7.353	24.244	15	.061	0

(2) 表 11.15 是模型的拟合参数表, 由于本例进行了差分运算, 因此常数项应该为 0, 模型的最终形式为

$$\nabla \nabla_{12} x_t = \frac{(1 - 0.999B)(1 - B^{12})}{1 - 0.588B} a_t$$

表 11.15 ARIMA 模型参数

				估算	标准误差	t	显著性
出口量-模型_1	出口量	不转换	常量	.724	.185	3.913	.000
			AR 延迟 1	.588	.101	5.795	.000
			差异	1			
			MA 延迟 1	.999	.836	1.194	.235
			AR, 季节性 延迟 1	-.332	.095	-3.493	.001
			季节性差异	1			

进一步简化得到：

$$\nabla \nabla_{12} x_t = \frac{(1 - 0.999x_{t-1})(1 - x_{t-12})}{1 - 0.588x_{t-1}} a_t$$

(3) 图 11-23 画出了残差的自相关函数和偏自相关函数，从图中可以看出，残差自相关函数和偏自相关函数都近似 0 阶截尾，说明残差是一个近似的白噪声序列，这也说明序列中的相关性信息都被提取完全，剩下的都是不相关的序列了。

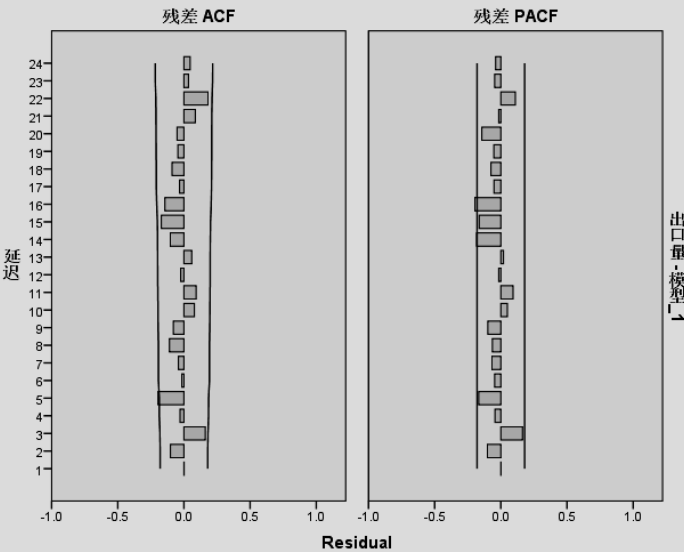


图 11-23 模型残差自相关函数和偏自相关函数图

(4) 最后通过序列图将“出口量”及预测值做模型的拟合效果图，图中的实线表示实际值，虚线表示预测值。从图 11-24 可以看出模型的拟合效果非常好，拟合值和观测值几乎重合。

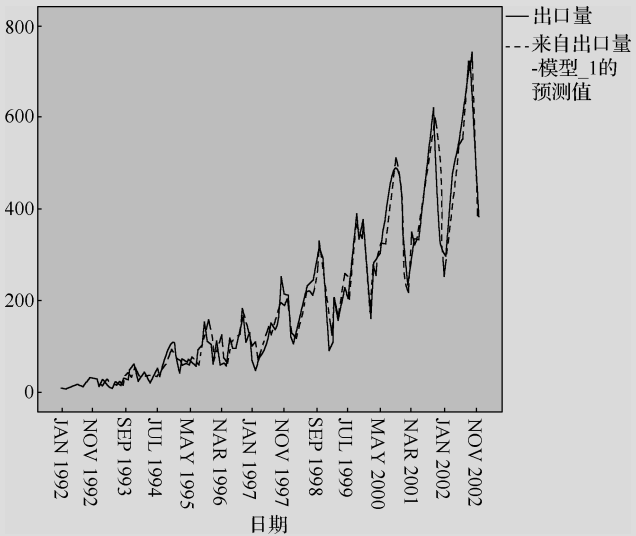


图 11-24 模型拟合值和观测值序列图

☆说明☆

- (1) 除了用指数平滑法和 ARIMA 模型进行建模外, SPSS 23 还提供了“专家建模器”方法, 专家建模器会自动查找每个相依序列的最佳拟合模型, 专家建模器既考虑指数平滑法模型也考虑 ARIMA 模型。例 11-5 如果采用“专家建模器”进行自动建模方法, 从运行的结果看系统采用了“指数平滑”法的“Winters 乘法”模型, 其决定系数(R 方)为 0.96, 正态化的 BIC 值为 7.255。比例 11-5 分析用的 ARIMA(1, 1, 1, 1, 1, 0)模型拟合效果略好, 这个过程请读者自己进行。
- (2) 在 SPSS 23 中进行预测时, 可以将自己创建好的模型保存下来, 以后直接用该模型进行预测, 其过程为“分析→时间序列预测→应用因果模型(应用传统模型)”, 请读者自己练习。

11.4 时间序列的季节性分解

11.4.1 基本概念及统计原理

在实际工作中, 人们经常按月(或年、季度、小时等)记录资料, 如每个月的出生人口数、死亡率、某种疾病的发病率、某产品的销售额等, 这些资料可能符合某种季节性分布, 但这些数值的大小往往受多种因素的影响, 从原始数据中很难看出季节趋势。

季节性分解法将时间序列分解为 3 个组成部分, 或称 3 个分量, 即“趋势分量”、“季节分量”和“随机波动”, 趋势分量采用多项式拟合, 季节分量用傅里叶变换估计, 其数学表达式为

$$Y_t = f(T_t, S_t, I_t) \quad (11.16)$$

式中, T_t 代表长期趋势(可以是线性趋势, 也可以是周期性波动或长周波动), S_t 为季节因子(幅度和周期固定的波动, 日历效应为常见的季节因子), I_t 为随机波动(可视为误差)。

常见的时间序列分解模型有加法和乘法两种, 加法模型为 $Y_t = T_t + S_t + I_t$, 乘法模型为 $Y_t = T_t \times S_t \times I_t$ 。相对而言, 乘法模型比加法模型用得更多。在乘法模型中, 时间序列值和长期趋势用绝对值表示, 季节变动和不规则变动用相对值(百分数)表示。

季节性分解要求无缺失数据, 在处理前数据已经定义好日期变量并指定周期。

11.4.2 季节性分解的实例分析

【例 11-6】对表 11.1 所示某企业的销售数据进行季节性分解。(参见数据文件: data11-1.sav.)

第 1 步 数据组织。

如例 11-1, 进行数据组织, 并定义“年份、月份”格式的日期变量。

第 2 步 观察数据序列的性质。

如例 11-3, 对销售额作时序图, 具体见图 11-5。从该时序图可以看出, 销售额总的趋势是增长的, 但增长并不是单调上升的, 而是有涨有落。这种升降不是杂乱无章的, 和季节或月份的季节因素有关。当然, 除了增长的趋势和季节影响之外, 还有些无规律的随机因素的作用。

第 3 步 季节性分析设置。

(1) 按“分析→时间序列预测→季节性分解”顺序打开“季节性分解”对话框, 并按如图 11-25 所示进行设置。

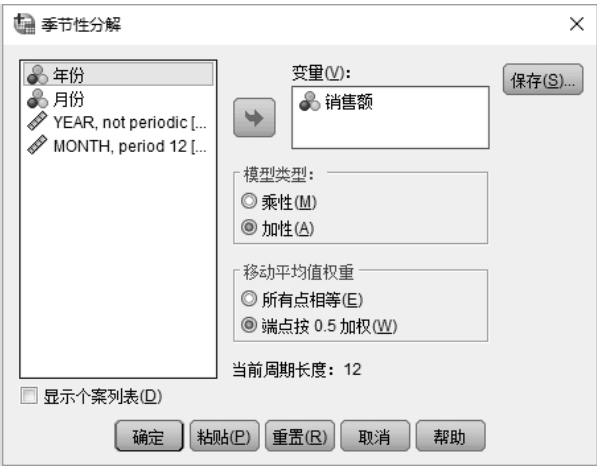


图 11-25 “季节性分解”对话框

现将对话框中的各项解释如下。

① “模型类型”选项组：季节性分解提供了用于对季节性因子建模的两种不同方法——乘法或加法。

- “乘性”模型：季节性成分是一个因子，用来与经过季节性调整的序列相乘以得到原始序列。实际上，“趋势”会评估与序列的总体水平成正比的季节性成分。无季节性变动的观察值的季节性成分为 1。
- “加性”模型：将季节性调整项加到季节性调整的序列以获取观察值。此调整尝试从序列中移去季节性影响，以查看可能被季节性成分“掩盖”的其他兴趣特征。实际上，“趋势”会评估不依赖于序列的总体水平的季节性成分。无季节性变动的观察值的季节性成分为 0。

② “移动平均值权重”选项组：允许指定在计算移动平均数时如何处理序列。这些选项仅在序列的周期为偶数时才可用。如果周期为奇数，则所有点的权重都相等。

- “所有点相等”选项：使用等于周期的跨度以及所有权重相等的点来计算移动平均数。如果周期是奇数，则始终使用此方法，此选项也是默认选项。
- “端点按 0.5 加权”选项：使用等于周期加 1 的跨度以及以 0.5 加权的跨度的端点计算具有偶数周期的序列的移动平均。

(2) “保存”对话框的设置：在图 11-25 上单击“保存...”按钮，打开“保存”对话框并做如图 11-26 所示的设置。其中有将季节分解后所创建的序列“添加到文件”、“替换现有项”和“不创建”3 项，本例选择“添加到文件”。

第 4 步 主要结果及分析。

主要结果如表 11.16、表 11.17 及图 11-27、图 11-28 所示，具体分析如下。

- (1) 表 11.16 是模型的描述表，显示了模型的名称、类型及季节性期间的长度等信息。
- (2) 表 11.17 是季节性因素表，由于季节性的影响，各月份的销售额有很大不同，可看出 11 月、12 月、3~5 月的季节性因子为负值，这几个月的销售情况比较差，12 月最差。同理，8 月份的销售情况最好。

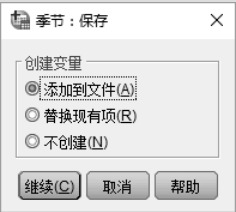


图 11-26 “季节：保存”对话框

表 11.16 模型描述表

模型名称	MOD_9	
模型类型	加性	
序列名称	1	销售额
季节性周期长度	12	
移动平均值的计算方法	跨度等于周期长度加 1，且端点按 0.5 加权	

正在应用来自 MOD_9 的模型指定项

(3) 图 11-27 是数据文件的数据视图。从该图中可以看到，数据文件中增加了 4 个序列：ERR_1 表示“销售额”序列进行季节性分解后的不规则或随机波动序列；SAS_1 表示“销售额”序列进行季节性分解除去季节性因素后的序列；SAF_1 表示“销售额”序列进行季节性分解产生的季节性因素序列；STC_1 表示“销售额”序列进行季节性分解出来的序列趋势和循环成分。

(4) 用数据文件中新增加的 4 个数据序列作序列图，如图 11-28 所示。

表 11.17 季节因子

序列名称: 销售额	
周期	季节因子
1	.97223
2	4.07407
3	-3.02840
4	-3.56468
5	-3.24368
6	1.90900
7	2.71091
8	4.44257
9	1.25785
10	2.13070
11	-1.47389
12	-6.18666

年份	月份	销售额	YEAR	MONT...	DATE	ERR_1	SAS_1	SAF_1	STC_1
1998	1	39.01	1998	1	JAN 1998	-1.72451	38.03777	.97223	39.76228
1998	2	44.24	1998	2	FEB 1998	-.07810	40.16593	4.07407	40.24404
1998	3	39.50	1998	3	MAR 1998	1.32086	42.52840	-3.02840	41.20754
1998	4	38.25	1998	4	APR 1998	.09522	41.81468	-3.56468	41.71946
1998	5	38.04	1998	5	MAY 1998	-.41631	41.28368	-3.24368	41.69999
1998	6	44.15	1998	6	JUN 1998	.58204	42.24100	1.90900	41.65896
1998	7	43.52	1998	7	JUL 1998	-.70757	40.80909	2.71091	41.51666
1998	8	46.65	1998	8	AUG 1998	.84020	42.20743	4.44257	41.36723
1998	9	42.30	1998	9	SEP 1998	.06183	41.04215	1.25785	40.98032
1998	10	41.87	1998	10	OCT 1998	-1.02337	39.73930	2.13070	40.76267
1998	11	39.52	1998	11	NOV 1998	-.06317	40.99389	-1.47389	41.05706
1998	12	35.18	1998	12	DEC 1998	.06349	41.36666	-6.18666	41.30317

图 11-27 数据文件中所创建的数据序列（部分）

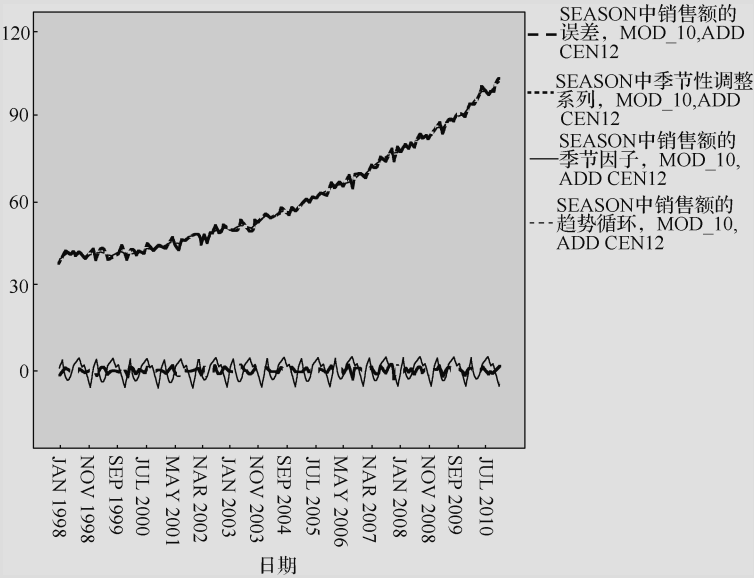


图 11-28 季节性分解后的各序列图

11.5 典型 案 例

11.5.1 中国社会消费品零售总额分析

改革开放以来，随着中国社会经济的快速发展，城乡居民和社会集团的消费水平不断提高，而且由于社会主义市场经济体制的建立，国内消费需求对经济增长所发挥的作用也更趋明显。为了深刻分析改革开放以来中国城乡居民和社会集团消费需求的发展态势，预测未来中国城乡居民和社会集团消费需求的基本走势，需要对中国国内消费需求的发展变化做数量分析。在各类与消费有关的统计数据中，社会消费品零售总额是表现国内消费需求最直接的数据，它反映各行业通过多种商品流通渠道向居民和社会集团供应的生活消费品总量，是研究国内零售市场变动情况、反映经济景气程度的重要指标。表 11.18 是 1978~2003 年中国社会消费品零售总额的月度数据。试分析改革开放以来中国社会消费品零售总额发展变化的基本趋势，分析是否存在季节变动或周期性变动规律，并对 2004~2007 年的社会消费品零售总额进行预测。（数据来源：袁卫 等，《统计学》（第二版），高等教育出版社；参见数据文件：data11-5.sav。）

表 11.18 中国 1978~2003 年社会消费品零售总额（单位：亿元）

年份	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
1978	134.3	119.4	128.3	126.4	128.8	127.8	121.1	118.4	125.7	123.6	128.5	145.2
1979	164.7	126.2	143.7	143.7	145.5	143.7	138.4	136.7	145.5	150.7149	149	164.7
1980	190.3	174.9	163.2	168.4	168.6	168.2	163.5	161.6	172.9	166.5	175.2	197.7
1981	212.1	177.9	182.9	184.2	184	182.4	175.6	172	184.9	184.7	195.1	224.8
1982	233.6	182	206.6	202.2	201.7	202.6	192.8	186.2	199.3	198.2	205.8	248.2
1983	243.2	217.5	226.2	223.5	221	220.5	205.8	206.9	218.8	216	235	282
1984	268.4	227.6	248.6	247	249.9	253.1	245.5	249.6	272.3	278.7	299.4	366.3
1985	338.9	320.4	308.5	290.1	291.8	295.2	282	284.2	317.8	319.5	341.7	411.1
1986	381.6	346	335.3	329.4	332.9	340.2	326.6	333.9	378.3	382	405.4	481.2
1987	445.2	394.1	391.5	383.9	389.7	400.7	390.7	396.1	439.3	451	471.2	561.6
1988	528.2	522.7	487.9	481	489.1	521.2	514.7	558.4	600.8	581.2	585.7	663.7
1989	638.5	615.6	609.6	581.1	575.6	576.2	546.2	546.1	586.5	571.8	580.3	665.8
1990	654.3	574.6	574.8	559.8	570.9	571.3	552.2	560.4	609.6	616.1	640.6	735.3
1991	709.7	705.1	644.3	650.7	635.9	645.7	624.2	634.5	694.8	706.6	736	831.6
1992	811.3	811.4	772.4	719	729	745.4	728.7	741.6	816.2	836.2	868.8	1028.6
1993	997.5	892.5	942.3	941.3	962.2	1006	963.8	959.8	1023	1051	1102	1416
1994	1192.2	1162.7	1166.1	1176.6	1213.7	1238.6	1251.5	1286	1399	1444.1	1553.8	1992.3
1995	1602.2	1491.5	1553.3	1548.7	1585.4	1639.7	1623.6	1637.1	1756	1818	1935.2	2389.5
1996	1909.1	1911.2	1860.1	1854.8	1898.3	1966	1888.7	1916.4	2083.5	2148.3	2290.1	2848.6
1997	2288.5	2213.5	2130.9	2100.5	2108.2	2164.7	2102.5	2104.4	2239.6	2348	2454.9	2881.7
1998	2549.5	2306.4	2279.7	2252.7	2265.2	2326	2286.1	2314.6	2443.1	2536	2652.2	3131.4
1999	2662.1	2538.4	2403.1	2356.8	2364	2428.8	2380.3	2410.9	2604.3	2743.9	2859	3383
2000	2962.9	2804.9	2626.4	2571.5	2636.9	2645.2	2596.9	2636.3	2854.3	3029.3	3107.8	3680
2001	3332.8	3047.1	2876.1	2820.9	2929.6	2908.7	2851.4	2889.4	3136.9	3347.3	3421.7	4033.3
2002	3596.1	3416	3197.4	3163.3	3320.5	3302.8	3244.2	3284.4	3627.2	3815.2	3831.1	4270.2
2003	4087.1	3706.4	3494.8	3406.9	3463.3	3576.9	3562.1	3609.6	3971.8	4204.4	4202.7	4735.7

案例分析：这是一个典型的时间序列问题，题目已对背景分析得比较清楚。由于涉及预测问

题，可以采用指数平滑法、ARIMA 模型等进行建模，然后预测。为了研究社会消费品零售总额是否存在季节性变动和周期性变动规律，可采用季节分解法对原数据序列进行分解后观察。

11.5.2 中国彩电出口数据分析

为了研究我国彩电出口的情况，某研究机构收集了从 1992~2002 年我国彩电出口的月度数据，如表 11.19 所示，试对这些年间我国彩电出口情况进行分析，主要研究以下几个问题：彩电出口的趋势如何？是否有季节性或周期性影响因素？并对 2003 年彩电出口数据进行预测。（参见数据文件：data11-6.sav。）

表 11.19 我国 1992~2002 年彩电出口数据（单位：万台）

	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
1992	12.53	13.73	24.45	28.75	32.45	31.11	25.94	32.98	43.49	42.94	63.29	77.28
1993	30.01	39.63	29.77	42.74	32.25	31.94	32.27	32.59	32.92	30.98	47.44	52.82
1994	24.08	16.42	31.24	29.33	31.88	30.09	28.08	32.99	44.99	47.57	50.36	75.19
1995	39.02	25.81	43.38	37.34	39.22	39.87	51.1	50.99	55.16	62.78	57.75	72.2
1996	28.76	39.38	46.1	39.41	38.74	40.18	45.59	43.31	46.68	54.17	53.65	61.12
1997	28.87	21.23	35.82	26.97	32.33	24.53	29.39	31.96	38.22	39.24	52.95	68.41
1998	29.99	37.09	37.7	35.33	29.53	53.64	28.95	25.88	37.61	39.83	28.44	54.85
1999	55.77	13.96	43.5	32.96	32.91	47.65	39.74	39.48	50.7	60.53	68.22	83.47
2000	66.35	70.35	86.19	87.5	61.19	93.23	89.31	88.37	90.05	90.06	107.56	101.63
2001	78.31	91.97	91.73	101.67	77.6	87.64	98.82	79.9	110.86	113.29	125.58	120.24
2002	101.65	93.53	127.04	133.68	143.76	155.5	170.59	168.96	186.16	181.91	253.78	201.14

案例分析：彩电出口数据是一个时间序列数据，题目给出了 11 年间按月份的出口数据，首先可通过作时序图观察数据序列的趋势，再用指数平滑法、ARIMA 模型或专家建模法进行建模以拟合彩电出口数据序列，并进行相关的预测。如果根据时序图观察数据序列存在季节性或周期性变动，则需用季节分解法对数据序列进行分解。

11.5.3 城市温度的季节性分解

为了研究某个城市的气温变化情况，某研究机构记录了 2005~2007 年该城市的月度平均气温，如表 11.20 所示。利用季节性分解法对该城市气温进行分析。（参见数据文件：data11-7.sav。）

表 11.20 某城市 2005~2007 年月平均气温（单位：℃）

年份	月份	温度	年份	月份	温度	年份	月份	温度
2005	1	-0.7	2006	1	-2.2	2007	1	-3.8
2005	2	2.1	2006	2	-0.4	2007	2	1.3
2005	3	7.7	2006	3	6.2	2007	3	8.7
2005	4	14.7	2006	4	14.3	2007	4	14.5
2005	5	19.8	2006	5	21.6	2007	5	20
2005	6	24.3	2006	6	25.4	2007	6	24.6
2005	7	25.9	2006	7	25.5	2007	7	28.2
2005	8	25.4	2006	8	23.9	2007	8	26.6
2005	9	19	2006	9	20.7	2007	9	18.6
2005	10	14.5	2006	10	12.8	2007	10	14
2005	11	7.7	2006	11	4.2	2007	11	5.4
			2006	12	0.9			

续表					
年份	月份	温度	年份	月份	温度
2008	1	-3.9	2009	1	-1.6
2008	2	2.4	2009	2	2.2
2008	3	7.6	2009	3	4.8
2008	4	15	2009	4	14.4
2008	5	19.9	2009	5	19.5
2008	6	23.6	2009	6	25.4
2008	7	26.5	2009	7	28.1
2008	8	25.1	2009	8	25.6
2008	9	22.2	2009	9	20.9
2008	10	14.8	2009	10	13
2008	11	4	2009	11	5.9
2008	12	0.1			

案例分析：气温的变化肯定会受季节性因素影响，主要通过季节性分解来处理，另外还要看该城市的气温变化是否有某种长期趋势及周期性变化等。

11.6 思考与练习

1. 对数据进行时间序列分析前，应做哪些准备工作？
2. 时间序列分析是建立在序列平稳条件上的，如何判断序列是否平衡？
3. 表 11.21 是 1992~2011 年我国煤矿事故死亡人数的统计数据，请用指数平滑法预测以后年度的煤矿事故死亡人数（数据来源：朱庆明等，《三次指数平滑法在煤矿事故预测中的应用研究》，中国安全生产科学技术；data11-8.sav。）

表 11.21 1992~2011 年我国煤矿事故死亡人数

年份	死亡人数	年份	死亡人数
1992	4942	2002	6995
1993	5283	2003	6434
1994	7016	2004	6027
1995	6387	2005	5896
1996	6404	2006	4746
1997	6753	2007	3786
1998	6134	2008	3218
1999	5518	2009	2631
2000	5798	2010	2433
2001	6850	2011	1973

4. 表 11.22 是某市连续 60 天（从 1999 年 5 月 19 日~1999 年 7 月 17 日）大气污染物总悬浮颗粒（TSP）的日均值监测结果，试对其建立 ARIMA 模型。（参见数据文件：data11-9.sav。）

表 11.22 某市连续 60 天大气污染情况数据

日期	TSP	日期	TSP	日期	TSP
19-May-99	0.175	8-Jun-99	0.087	28-Jun-99	0.123
20-May-99	0.191	9-Jun-99	0.16	29-Jun-99	0.135
21-May-99	0.185	10-Jun-99	0.158	30-Jun-99	0.161

续表					
日期	TSP	日期	TSP	日期	TSP
23-May-99	0.13	12-Jun-99	0.153	2-Jul-99	0.134
24-May-99	0.147	13-Jun-99	0.112	3-Jul-99	0.129
25-May-99	0.138	14-Jun-99	0.111	4-Jul-99	0.12
26-May-99	0.167	15-Jun-99	0.109	5-Jul-99	0.11
27-May-99	0.178	16-Jun-99	0.147	6-Jul-99	0.12
28-May-99	0.193	17-Jun-99	0.155	7-Jul-99	0.161
29-May-99	0.202	18-Jun-99	0.144	8-Jul-99	0.139
30-May-99	0.138	19-Jun-99	0.158	9-Jul-99	0.197
31-May-99	0.186	20-Jun-99	0.14	10-Jul-99	0.158
1-Jun-99	0.15	21-Jun-99	0.14	11-Jul-99	0.184
2-Jun-99	0.151	22-Jun-99	0.114	12-Jul-99	0.136
3-Jun-99	0.168	23-Jun-99	0.14	13-Jul-99	0.149
4-Jun-99	0.226	24-Jun-99	0.153	14-Jul-99	0.208
5-Jun-99	0.197	25-Jun-99	0.13	15-Jul-99	0.163
6-Jun-99	0.154	26-Jun-99	0.15	16-Jul-99	0.16
7-Jun-99	0.074	27-Jun-99	0.12	17-Jul-99	0.137

（提示：大气污染一般会受工作日的影 响，可能会在一星期内呈周期性波动，故可以按“星期、日”的格式定义日期变量，1999 年 5 月 19 日为星期三，故在对第一个个案的星期几处应输入 3。）

5. 记录中国某城市从 1990 年~2009 年的平均气温、平均最高气温、平均最低气温、平均相对湿度、月降水量、月日照时数、平均本站气压和菌痢率，其中菌痢率如表 11.23 所示。利用季节分析模型对该城市菌痢率进行分析。（数据来源：武松 等，《SPSS 统计分析大全》，清华大学出版社；参见数据文件：data11-10.sav。）

表 11.23 某城市 1990~2009 年每月菌痢率（%）

年份\月份	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
1990	1.67	1.57	2.34	3.56	5.42	7.66	8.43	12.15	12.07	10.08	5.77	3.48
1991	1.98	1.43	1.83	3.18	5.66	8.19	14.37	27.56	20.63	10.09	2.14	1.12
1992	2.42	1.54	1.49	2.31	3.34	4.45	5.11	7.27	6.63	3.37	1.23	0.36
1993	0.66	0.99	1.14	1.14	3.32	2.99	4.87	6.72	7.12	4.16	2.51	1.80
1994	1.45	1.85	1.78	1.83	2.83	6.55	10.60	22.36	32.81	27.88	10.95	3.00
1995	2.24	1.67	2.43	4.08	7.89	13.59	16.51	19.19	15.11	14.16	3.81	1.20
1996	2.68	1.89	2.83	3.50	6.65	8.56	14.48	15.44	16.31	12.54	4.06	1.86
1997	1.70	1.33	1.96	2.07	3.71	7.40	8.81	8.08	6.45	5.63	3.86	1.77
1998	1.04	1.23	1.46	2.09	3.11	4.07	6.67	9.19	7.57	7.62	3.96	1.90
1999	1.22	0.83	1.21	1.75	3.50	5.61	6.99	5.80	5.33	3.49	1.81	1.16
2000	0.90	1.02	1.32	1.75	2.91	5.37	6.03	5.22	4.35	3.50	1.59	1.05
2001	0.92	0.97	1.19	1.43	2.79	4.86	5.25	4.77	4.02	3.05	2.04	1.05
2002	1.19	1.01	1.29	1.59	2.37	4.14	5.23	4.94	3.82	2.35	1.20	0.85
2003	0.81	0.93	1.17	1.20	1.80	3.10	6.36	4.91	4.67	2.88	1.32	0.70
2004	0.68	0.79	0.88	1.33	2.75	4.74	5.05	4.46	3.86	2.52	1.35	0.71
2005	0.67	0.54	0.81	0.97	2.09	3.96	5.83	4.11	3.46	2.49	1.21	0.79
2006	0.54	0.49	0.68	0.88	1.62	3.06	4.56	3.72	2.75	2.81	1.13	0.61
2007	0.58	0.47	0.71	0.91	2.07	3.17	3.32	2.38	2.25	1.68	0.76	0.56
2008	0.48	0.43	0.55	0.84	1.60	2.19	2.20	2.09	1.90	1.65	0.75	0.53
2009	0.40	0.48	0.64	0.76	1.38	2.19	2.75	2.36	2.12	1.44	0.75	0.62

第 12 章 信 度 分 析

在分析问题时，我们常借助于量表或问卷进行。量表是否能测得所需的测量结果？测量结果的可靠性如何？需要对量表的效度、信度进行分析。

效度（Validity）是指量表是否真正反映了我们希望测量的问题。例如，智商测验是否真正反映了智力的高低？生存质量调查是否真正反映了人们的生存质量？抑郁量表调查是否真正反映了人们的抑郁程度？这些都是关于效度的问题。对于效度，我们没有绝对准确的答案。尽管不可能证明效度，但是可以用一些指标来评价效度。一般来说，有 4 种类型的效度：内容效度、实证效度、结构效度和区分效度。内容效度是一种基于概念的评价指标，其他三种效度是基于经验的评价指标。如果一个量表实际上是有效的，那么我们希望上述 4 种指标都比较满意。一般使用的效度评估方法，主要有判断法和实证法，前者着重于测量特性与质的评估，通常依赖研究者对数据的主观判断；实证法则根据具体客观的量化指标来进行效度的评估。进行实证效度评估的统计方法有：相关分析、多元回归分析、因子分析、结构方程等。

信度（Reliability Analysis）是指测量的一致性。例如，准备调查你的年收入，你第一次回答是 20000 元，如果能够消除你对问题和回答的记忆，过一段时间后问你同一个问题（当然也可以通过调查其他人了解你的年收入情况），通过考察你（你们）对同一问题的多次回答，可以判断答案的一致性如何。答案的波动越大，信度越低；回答的一致性越好，信度越高。信度本身与测量所得结果正确与否无关，它的功能在于检验测量本身是否稳定。制作完成一份量表或问卷后，首先应该对该量表进行信度分析，确保其可靠性和稳定性，以免影响问卷内容分析结果的准确性。

根据测试时间和测试内容，信度又可分为内在信度和外在信度。内在信度是指一组问题（也可以称为题项）是否测量同一个概念，即这些问题的内在一致性如何，能否稳定地衡量这一概念，最常用的检测方法是 Cronbach 系数；而外在信度是指对相同的测试者在不同时间测得的结果是否一致，重测信度是外在信度最常用的检验法。

效度与信度的关系是，信度为效度的必要而非充分条件，即有效度一定有信度，但有信度不一定有效度。本章只讨论信度问题，并就内在信度和外在信度等进行介绍。

12.1 内在信度分析

12.1.1 基本概念及统计原理

1. 基本概念

内在信度也称为内部一致性，用以衡量组成量表题项的内在一致性程度如何。常用的检测方法是 Cronbach α 系数法和折半（Split-half）系数法。

2. 统计原理

内在信度一般采用 Cronbach α （克隆巴赫）信度系数进行评价，也可采用折半系数法进行检测。

(1) Cronbach α 信度系数

Cronbach α 信度系数是目前最常用的信度系数。其公式为

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k S_i^2}{S_x^2} \right) \quad (12.1)$$

式中, k 为测验的题目数; S_i 为第 i 题得分数的方差; S_x 为测验总分之方差。Cronbach α 系数在 0~1 之间。从经验来看, 如果 $\alpha \geq 0.9$, 则认为量表的内在信度很高; 如果 $0.8 \leq \alpha < 0.9$, 则认为内在信度较好; 如果 $0.7 \leq \alpha < 0.8$, 则认为量表设计可以接受; 如果 $0.6 \leq \alpha < 0.7$, 则认为量表设计勉强可以接受, 最好进行适当修改; 如果 $0.5 \leq \alpha < 0.6$, 表明量表设计不理想, 需重新编制或修订; 如果 $\alpha < 0.5$, 表明量表非常不理想, 应舍弃。

(2) 折半信度系数

折半信度是在测试后对测试项目按奇项、偶项或其他标准分成两半, 分别记分, 由两半分数之间的相关系数得到信度系数。因此它实际上是检验一个测试内部一致性的粗略估计。折半信度是建立在相关系数基础上的, 但它在相关系数(由于折半, 所以相关系数只是半个测验的信度, 可能会低估原测验的信度)基础上, 需要进行斯皮尔曼-布朗公式校正, 校正公式为

$$r = \frac{2r_{xx}}{1+r_{xx}} \quad (12.2)$$

式中, r_{xx} 为两半测验分数的相关系数。

折半信度面临的主要问题是如何将问题分成两半。一般事实式的问题是不太容易折半的, 例如年龄和教育程度是无法相比的。因此这种方法一般不适合于事实式量表。态度式量表一般围绕某个主题进行多种正、反面的陈述, 由被调查者对陈述进行选择。例如“很不满意”、“不满意”、“既非满意也非不满意”、“满意”、“很满意”中的一个, 对以上 5 种选择分别赋予 1~5 分, 然后将该量表的全部题项分成尽可能相近的两半, 按前后两部分或按题号的奇偶性分都可以, 只是要注意两部分必须尽可能相当(内容、形式、题数等), 不同的折半法可能会得到不同的结果。

☆说明☆

- ◆ 折半信度系数也可以用于度量量表的外在信度。在用于计算量表外在信度时需将两个表复合, 然后求其折半信度系数, 这就得出了两张量表之间的一致性情况。

12.1.2 内在信度实例分析

【例 12-1】 在学生的性格特征调查中, 共选了 10 名学生在 8 个项目上进行测试, 其数据如表 12.1 所示, 试对其进行内在信度分析。(参见数据文件: data12-1.sav。)

第 1 步 分析。

本例通过求 Cronbach α 系数来衡量其内在一致性。

第 2 步 数据组织。

按如表 12.1 所示的表头定义变量, 输入数据并保存。

第 3 步 内在一致性分析的设置。

(1) 按“分析→标度→可靠性分析”顺序打开“可靠性分析”对话框, 将衡量性格的 8 个变量移入“项”框中, 如图 12-1 所示。

表 12.1 学生性格特征调查数据

序号	内向性	活动性	支配性	深思性	健壮性	稳定性	社会性	激动性
1	4	6	5	5	5	3	5	4
2	2	5	4	5	5	3	4	2
3	3	5	3	6	4	1	3	1
4	5	6	4	7	5	5	6	2
5	3	6	5	6	4	4	6	3
6	3	3	3	2	1	1	2	1
7	4	6	6	6	5	6	5	1
8	7	6	2	6	4	5	6	4
9	2	3	2	2	7	4	7	2
10	2	3	4	4	5	6	3	1



图 12-1 “可靠性分析”对话框

(2) 模型选择，本例选择使用 Cronbach's α 信度系数法，共有以下 5 种。

- “Alpha”模型：默认选项，计算量表内在一致性的 Cronbach α 系数。
- “折半”信度系数模型：计算折半信度系数，输出结果中给出 Guttman 和 Spearman-Brown 折半信度系数以及折半后两部分的 Cronbach α 系数，考察两部分之间的相关性。
- “格特曼”模型：计算最低下限的真实信度法，输出结果中产生 6 个信度系数，从 $\lambda_{bd1} \sim \lambda_{bd6}$ 。
- “平行”模型：计算各评估项目变异数同质时的最大概率（maximum-likelihood）信度，该模型假设所有项目具有相等的方差和相等的方差误差。
- “严格平行”模型：该模型是假设测试项目具有相等均值的平行模型法。输出结果中包含模型的拟合优度检验、各评估项目的方差估计值、项内相关系数、信度的无偏估计等统计量。

(3) “统计”对话框设置：单击图 12-1 上的“统计(S)…”按钮，打开“可靠性分析：统计”对话框，并按图 12-2 进行设置。

现对其中各选项解释如下。

① “描述”选项组：对各项目、维度得分情况和项目与量表总体特征的关系进行描述性统计。

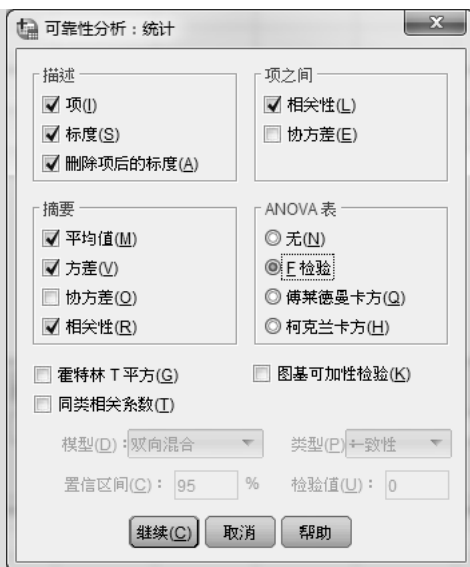


图 12-2 “可靠性分析：统计”对话框

- “项”选项：输出各评估项目的基本描述性统计量，包括项内均值和标准差等。
- “标度”选项：显示量表总体（或维度）的均值和方差。
- “删除项后的标度”选项：输出剔除某评估项目后的基本统计量，以便对各个评估项目逐个评价。

② “项之间”选项组：输出项内统计量。

- “相关性”选项：计算项目间的两两相关系数。
- “协方差”选项：计算项目间的两两协方差值。

③ “摘要”选项组：用于计算“描述性”选项组中指定对象的相关值，包括“平均值”、“方差”、“协方差”和“相关性”等。

④ “ANOVA 表”选项组：方差分析表，给出了用于检验同一评估对象在各评估项目上的得分是否具有一致性的方法，包括以下几个选项。

- “无”选项：不检验。
- “F 检验”选项：表示重复测量的方差分析，适合于数据为定距型且服从正态分布的情况。
- “傅莱德曼卡方”选项：表示进行非参数检验中的多配对样本 Friedman 检验，适合数据为非正态分布或定序型，计算 Friedman 值和 Kendall 一致性系数，在 ANOVA 表中，利用卡方检验代替 F 检验。
- “柯克兰卡方”选项：计算 Cochran 卡方值，适用于所有项目均为二分变量（0 和 1 的情况）。

⑤ “霍特林 T 平方”选项：计算 Hotelling T 平方值，检验所有评估项目的均值是否相等的多变量检验。

⑥ “图基可加性检验”选项：Tukey 检验，评估项目中是否存在倍增交互效应。

⑦ “同类相关系数”选项：进行一致性测度或个案数值的一致性检验。计算组内相关系数需选择计算方法和相关类型，选择该项时，以下两个选项变为可用。

“模型”下拉列表：用于选择计算组内相关系数的方法，包括以下 3 种模型。

- “双向混合”模型：当个案效应和评估项目效应均为固定时选择此项。

- “双向随机”模型：当个案效应和评估项目效应均为随机时选择此项。
- “单向随机”模型：当个案效应为随机时选择此项。
- “类型”下拉列表：用于选择指标类型，该下拉列表框有以下两项：“一致性”和“绝对一致性”。
- “置信区间”文本框：用于指定置信区间的水平，默认值为 95%。

“检验值”文本框：用于指定假设检验过程的检验值，默认值为 0，可输入 0~1 之间的数值，用于类间相关系数的比较。

第 4 步 主要结果及分析。

本例信度分析的主要结果如表 12.2~表 12.9 所示，具体分析如下。

- (1) 表 12.2 是个案的摘要情况表，可看出其中有 10 个个案参与信度分析，不含缺失值。
- (2) 表 12.3 是 Cronbach α 系数表，可知 α 系数为 0.790，其标准化后的 α 系数为 0.790，说明量表的信度是可以接受的，当然也有进一步优化的空间。

表 12.2 个案处理摘要

表 12.3 信度分析的 Cronbach 系数表

		个案数	%
个案	有效	10	100.0
	排除 ^a	0	.0
	总计	10	100.0

克隆巴赫 Alpha	基于标准化项的克隆巴赫 Alpha	项数
.790	.790	8

a. 基于过程中所有变量的成列删除。

- (3) 表 12.4 是所有评估项目的均值、标准差及个案数情况。

表 12.4 评估项目的基本描述

	平均值	标准差	个案数
内向性	3.50	1.581	10
活动性	4.90	1.370	10
支配性	3.80	1.317	10
深思性	4.90	1.729	10
健壮性	4.50	1.509	10
稳定性	3.80	1.814	10
社会性	4.70	1.636	10
激动性	2.10	1.197	10

- (4) 表 12.5 是评估项目间的相关系数矩阵，可看出“活动性”和“深思性”的相关性最大，为 0.886，“社会性”和“支配性”的相关性最小，绝对值为 0.031。

表 12.5 评估项目的相关系数矩阵

	内向性	活动性	支配性	深思性	健壮性	稳定性	社会性	激动性
内向性	1.000	.641	-.107	.549	-.163	.271	.365	.558
活动性	.641	1.000	.480	.886	.081	.215	.431	.549
支配性	-.107	.480	1.000	.430	.056	.307	-.031	-.056
深思性	.549	.886	.430	1.000	.106	.312	.263	.274
健壮性	-.163	.081	.056	.106	1.000	.528	.652	.154
稳定性	.271	.215	.307	.312	.528	1.000	.502	.113
社会性	.365	.431	-.031	.263	.652	.502	1.000	.584
激动性	.558	.549	-.056	.274	.154	.113	.584	1.000

(5) 表 12.6 是评估项目的总体描述性情况表，显示了 10 个学生在 8 个评估项目上的均值、方差和项之间相关性的基本描述，包括平均值、最小值、最大值、方差等。

表 12.6 评估项目的总体描述性情况表

	平均值	最小值	最大值	全距	最大值/最小值	方差	项数
项平均值	4.025	2.100	4.900	2.800	2.333	.899	8
项方差	2.347	1.433	3.289	1.856	2.295	.410	8
项间相关性	.320	-.163	.886	1.049	-5.439	.067	8

(6) 表 12.7 是对所有评估项目的描述性情况表，显示了将某一项从量表中删除的情况下，量表的平均分、方差、每个项目得分与剩余各项目得分之间的相关系数，以该项目为自变量，所有其他项目为因变量建立回归方程的 R 方值以及 Cronbach α 值。从表中可以看出，“活动性”与其他项目之间的相关性最高，为 0.752，而且“活动性”与其他项目的复相关系数（R 方）也最高，为 0.982，这表明“活动性”与其他项目的关系最为密切。同时也可以看出，如果删除“支配性”，则其 α 系数变成了 0.802，有所提升，但幅度并不大。

表 12.7 所有评估项目的描述性情况表

	删除项后的 标度平均值	删除项后的 标度方差	修正后的 项与总计相关性	平方多重相关性	删除项后的 克隆巴赫Alpha
内向性	28.70	48.233	.460	.905	.773
活动性	27.30	45.122	.752	.982	.730
支配性	28.40	54.489	.238	.908	.802
深思性	27.30	43.567	.626	.946	.744
健壮性	27.70	51.567	.323	.752	.793
稳定性	28.40	45.156	.509	.879	.766
社会性	27.50	44.500	.626	.850	.745
激活性	30.10	51.211	.479	.701	.772

(7) 表 12.8 是所有项目的描述表，可看出 8 个项目的总分均值、方差、标准偏差。

表 12.8 所有项目的描述表

均值	方差	标准偏差	项数
32.20	60.844	7.800	8

(8) 表 12.9 是方差分析表。 $F=5.635$ ，显著性概率值为 0.000，因此应拒绝 F 检验的零假设，认为各项目的均值总体上存在显著性差异，各项得分不全部相等，即有一些项与其他项目存在不一致和不相关性。

表 12.9 ANOVA 分析表

		平方和	自由度	均方	F	显著性
人员间		68.450	9	7.606		
人员内	项间	62.950	7	8.993	5.635	.000
	残差	100.550	63	1.596		
	总计	163.500	70	2.336		
总计		231.950	79	2.936		

总平均值 = 4.03

综上所述，该量表的信度不是太高，应做相应调整。

【例 12-2】 对学生的性格特征数据（data12-1.sav）用折半法求信度系数。

具体过程和设置与例 12-1 类似，主要结果如表 12.10～表 12.12 所示（与前面相类似的表格不再逐一列出）。

（1）表 12.10 中将 8 个项目分成两部分，第一部分（a 项）包含“内向性”、“活动性”、“支配性”、“深思性”，第二部分（b 项）包含“健壮性”、“稳定性”、“社会性”、“激动性”。从表中可以看出，两部分的相关系数为 0.370，相关程度较低，表明该量表需重新编制或修订。各部分的信度系数一般，第一部分为 0.790，第二部分为 0.749，说明它们内部各自的一致性也不高。表 12.10 还分别给出了斯皮尔曼-布朗（Spearman-Brown）系数和格特曼折半（Guttman Split-Half）系数，均很小，也说明量表设计不够科学或所调查的数据不够准确，需进一步调整。

表 12.10 折半信度的信度系数统计表

克隆巴赫 Alpha	第一部分	值	.790
		项数	4 ^a
	第二部分	值	.749
		项数	4 ^b
	总项数		8
形态之间的相关性			.370
斯皮尔曼-布朗系数	等长		.540
	不等长		.540
格特曼折半系数			.540

a. 项为：内向性，活动性，支配性，深思性。

b. 项为：健壮性，稳定性，社会性，激动性。

（2）表 12.11 是分半后的项目描述性表，它显示了两部分在 4 个项目和总项目上的平均值、方差、相关系数的基本描述。

表 12.11 折半后的项目描述性表

		平均值	最小值	最大值	全距	最大值 / 最小值	方差	项数
项平均值	第一部分	4.275	3.500	4.900	1.400	1.400	.536	4 ^a
	第二部分	3.775	2.100	4.700	2.600	2.238	1.396	4 ^b
	两部分	4.025	2.100	4.900	2.800	2.333	.899	8
项方差	第一部分	2.275	1.733	2.989	1.256	1.724	.337	4 ^a
	第二部分	2.419	1.433	3.289	1.856	2.295	.605	4 ^b
	两部分	2.347	1.433	3.289	1.856	2.295	.410	8
项间相关性	第一部分	.480	-.107	.886	.993	-8.304	.099	4 ^a
	第二部分	.422	.113	.652	.540	5.794	.048	4 ^b
	两部分	.320	-.163	.886	1.049	-5.439	.067	8

a. 项为：内向性，活动性，支配性，深思性。

b. 项为：健壮性，稳定性，社会性，激动性。

（3）表 12.12 是两部分量表的描述性情况表，显示两分量表总分的平均值、方差和标准差情况。

表 12.12 两分量表的描述性情况表

	平均值	方差	标准差	项数
第一部分	17.10	22.322	4.725	4 ^a
第二部分	15.10	22.100	4.701	4 ^b
两部分	32.20	60.844	7.800	8

a. 项为：内向性，活动性，支配性，深思性。

b. 项为：健壮性，稳定性，社会性，激动性。

12.2 再测信度分析

12.2.1 基本概念及统计原理

1. 基本概念

同一个测验项目，对同一组人员进行前后两次测试，两次测试所得分数的相关系数即再测信度。它反映两次测验结果有无变动，也就是测验分数的稳定程度，故又称为稳定性系数。

2. 统计原理

(1) Pearson 相关系数

再测信度实质是求同一量表在两次测试中的相关系数，通常是求式（12.3）所示的 Pearson 相关系数。

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (12.3)$$

再测信度必须满足以下几个假设：

- ① 所测量的特质必须是稳定的；
- ② 遗忘和练习的效果相同；
- ③ 两次测试期间被试者对问题的熟悉情况没有差别。

若以上 3 条不易做到，则不宜采用再测信度。在获得再测信度时，两次测量的时间间隔要适当。时间太短会增加稳定性，时间太长容易受到其他因素变化的影响。同时第二次测量时，需要调动被试者的积极性，避免被试者对于重复测量引起不必要的情绪反应，影响第二次测试的效果。

(2) Cohen Kappa 指数

Kappa 指数用来描述两种测量手段的一致性，如果其中一种手段为标准测量手段，那么它就是标准度。

当 χ^2 检验认为两种测量结果具有一致性后，可以进一步计算反映一致性的指标 Kappa 指数。具体公式如下：

$$\text{符合率： } P_0 = \frac{\sum A_{ii}}{n} \quad (12.4)$$

$$\text{不一致率： } Q_0 = 1 - P_0 \quad (12.5)$$

$$\text{期望符合率: } P_e = \frac{\sum m_i m_i}{n^2}$$

(12.6)

$$\text{则 Kappa 指数: } \kappa = \frac{P_0 - P_e}{1 - P_e}$$

(12.7)

我们用这个 Kappa 指数来描述两种测量手段的一致性。根据经验, Kappa > 0.75, 可以认为一致性较好; 0.4≤Kappa ≤0.75, 说明一致性中等; Kappa < 0.4, 则表明一致性较差。

12.2.2 再测信度实例分析

【例 12-3】 心理调查第一次调查的数据如表 12.1 所示, 第二次调查的数据如表 12.13 所示。试对该量表进行再测信度分析。(参见数据文件: data12-2.sav。)

表 12.13 第二次性格调查数据

序号	内向性 1	活动性 1	支配性 1	深思性 1	健壮性 1	稳定性 1	社会性 1	激动性 1
1	3	5	6	5	4	4	4	4
2	2	5	5	5	3	4	5	3
3	3	5	3	6	5	2	3	2
4	4	6	4	7	5	4	5	3
5	3	6	5	6	4	4	4	5
6	4	3	2	2	1	1	3	2
7	4	6	6	6	4	5	5	2
8	6	6	2	5	4	5	5	4
9	3	3	3	3	6	5	6	3
10	2	3	4	4	5	6	4	2

第 1 步 分析。

进行再测信度分析。

第 2 步 数据组织。

建立“内向性”~“激动性”8 个变量及这 8 个变量的总分“total”(总分通过“转换→变量计算”来计算)变量, 和“内向性 1”~“激动性 1”及这 8 个变量的部分“total1”外加一个“序号”变量, 共 19 个变量, 如图 12-3 所示。

序	内向性	活动性	支配性	深思性	健壮性	稳定性	社会性	激动性	total	内向性1	活动性1	支配性1	深思性1	健壮性1	稳定性1	社会性1	激动性1	total1
1	1	4	6	5	5	3	5	4	37	3	5	6	5	4	4	4	4	35
2	2	2	5	4	5	5	3	4	30	2	5	5	5	3	4	5	3	32
3	3	3	5	3	6	4	1	3	26	3	5	3	6	5	2	3	2	29
4	4	5	6	4	7	5	6	2	40	4	6	4	7	5	4	5	3	38
5	5	3	6	5	6	4	4	6	37	3	6	5	6	4	4	4	5	37
6	6	3	3	3	2	1	1	2	16	4	3	2	2	1	1	3	2	18
7	7	4	6	6	6	5	6	5	39	4	6	6	6	4	5	5	2	38
8	8	7	6	2	6	4	5	6	40	6	6	2	5	4	5	5	4	37
9	9	2	3	2	2	7	4	7	29	3	3	3	3	6	5	6	3	32
10	10	2	3	4	4	5	6	3	28	2	3	4	4	5	6	4	2	30

图 12-3 数据组织图

第 3 步 再测信度分析设置。

按“分析→相关→双变量”顺序打开相关性对话框, 将前后两次调查的变量及数据移入“变量”对话框, 在“相关系数”框中选择“皮尔逊”系数, 即求 Pearson 相关系数, 在“显著性检验”框中选择“双尾”, 进行双尾检测, 之后提交系统运行。

第 4 步 主要结果及分析。

运行后的相关系数如表 12.14 所示（由于输出的原始表格太复杂，对输出的结果表进行了整理），从表中可以看出各变量的相关系数比较高，其中总分的相关系数为 0.976，说明量表的再测信度很好。

表 12.14 各项目的再测信度（两次的相关性系数）

内向性	活动性	支配性	深思性	健壮性	稳定性	社会性	激动性	总分
0.898	0.973	0.906	0.966	0.833	0.904	0.787	0.880	0.976

【例 12-4】两名放射科医师对 200 名肺病可疑患者的 X 光片进行读片的诊断结果如表 12.15 所示（其中“0”表示正常，“1”表示病重为一级，“2”表示病重为二级），请计算 Kappa 指数。（数据来源：宇传华，《SPSS 与统计分析》，电子工业出版社；参见数据文件：data12-3.sav。）

表 12.15 200 例肺病 X 光片诊断结果

第一次检查	第二次检查	例数
0	0	78
0	1	5
0	2	0
1	0	6
1	1	56
1	2	13
2	0	0
2	1	10
2	2	32

第 1 步 分析。

由于考察的是两个医师读 X 光片的一致性，故可用 Kappa 指数度量。

第 2 步 数据组织。

建立“第一次检查”、“第二次检查”和“例数”三个变量，输入数据并保存。

第 3 步 数据加权。

按“数据→个案加权”打开加权设置对话框，并将“例数”变量作为加权变量。

第 4 步 计算 Kappa 指数设置。

按“分析→描述统计→交叉表”打开“交叉表”对话框，并按图 12-4 进行设置。打开“统计”对话框，按图 12-5 进行设置，表示输出 χ^2 统计量和 Kappa 指数。



图 12-4 “交叉表”对话框

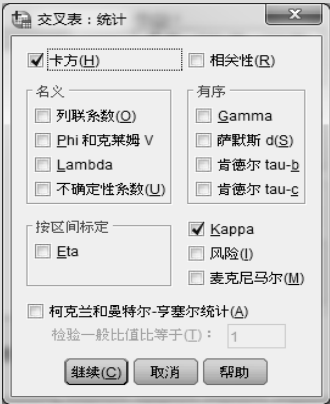


图 12-5 “交叉表：统计”对话框

第 5 步 主要结果及分析。

运行结果如表 12.16~表 12.18 所示，具体解释如下。

(1)表 12.16 是交叉列联表的情况,说明第一次检查和第二次检查相符合的记录数为 166(78 + 56 + 32 = 166), 占总数的 83% (166/200 × 100% = 83%)。

表 12.16 交叉列联表

		第二次检查			总计
		正常	一级	二级	
第一次检查	正常	78	5	0	83
	一级	6	56	13	75
	二级	0	10	32	42
总计		84	71	45	200

(2)表 12.17 是卡方检验结果表,给出了皮尔逊卡方检验结果,且对应的渐进显著性概率(双侧)为 0.000,说明第一次检查和第二次检查存在相关性。于是,进一步计算衡量两次检查结果一致性的 Kappa 指数。

表 12.17 卡方检验结果表

	值	自由度	渐进显著性(双侧)
皮尔逊卡方	219.384 ^a	4	.000
似然比	234.563	4	.000
线性关联	146.290	1	.000
有效个案数	200		

a. 0 个单元格 (0.0%) 的期望计数小于 5。最小期望计数为 9.45。

(3)表 12.18 给出了具体的 Kappa 指数为 0.737,显著性概率为 0.000,说明两次检查结果的一致性较好。

表 12.18 具体的 Kappa 指标表

		值	渐近标准化误差 ^a	近似 T ^b	渐进显著性
协议测量	Kappa	.737	.041	14.424	.000
有效个案数		200			

a. 未假定原假设。

b. 在假定原假设的情况下使用渐近标准误差。

12.3 评分者信度分析

12.3.1 基本概念及统计原理

1. 基本概念

所谓评分者信度 (Scorer Reliability),指的是多个评分者给同一批人进行评分的一致性程度。例如,在教育和心理测量中,常常关心不同的评分者对同一个主观题的评分是否一致;在医学临床疗效评价中,常常关心不同的医生对同一个患者的评价是否一致。当评分者人数为 2 时,可以采用 Pearson 或 Spearman 相关系数评价一致性;当评分者人数多于 2 时,可以采用肯德尔(Kendall)协同系数考察评分者信度。本节介绍肯德尔协同系数。

2. 统计原理

Kendall 协同系数的计算公式为

$$W = 12 \times \frac{R_i^2 - (\sum R_i)^2 / N}{K^2(N^3 - N)}$$
 (12.8)

式中， K 表示评分者人数； N 是被评分者人数； R_i 是第 i 个被评分者所得分数的水平等级之和。
若评分中出现相同的等级，则需要计算校正的系数。公式如下：

$$W = 12 \times \frac{R_i^2 - (\sum R_i)^2 / N}{K^2(N^3 - N) - K \sum \sum (n^3 - n) / 12}$$
 (12.9)

式中， n 为相同等级的个数。

12.3.2 评分者信度实例分析

【例 12-5】 三名神经内科医生对 6 名重症肌无力患者分别进行肌力的评分，结果如表 12.19 所示，试评价三名医生评价的一致性。（参见数据文件：data12-4.sav。）

第 1 步 分析。

这是一个考察几个人对同一批患者评价一致性的问题，考虑用 Kendall 协同系数来度量。

第 2 步 数据组织。

建立“序号”、“医生甲”、“医生乙”和“医生丙”四个变量，输入数据并保存。

第 3 步 Kendall 系数求解设置。

按“分析→非参数检验→旧对话框→K 个相关样本”顺序打开“针对多个关联样本的检验”对话框，并按图 12-6 进行设置，之后提交系统运行。

表 12.19 三名医生对肌无力患者的评价结果

序号	医生甲	医生乙	医生丙
1	35	32	25
2	40	36	30
3	37	31	28
4	30	30	24
5	38	35	31
6	42	40	32



图 12-6 “针对多个相关样本的检验”对话框

第 4 步 主要结果及分析。

运行结果如表 12.20 所示，不仅给出了肯德尔协同系数为 0.964，而且还给出了显著性水平为 0.003，说明三个医生评分结果具有较好的一致性。

表 12.20 检验统计结果表

个案数	6
肯德尔 W ^a	.964
卡方	11.565
自由度	2
渐近显著性	.003

a. 肯德尔协同系数

12.4 典型案例

12.4.1 Oxford 学习策略量表信度分析

某研究机构曾用 Oxford 学习策略量表对某校的 82 个大学生进行过问卷调查。该量表共分为 6 个子量表，如表 12.21 所示。现以该量表的调查数据为依托进行信度分析，探索该量表的内部一致性。（数据来源：冯岩松，《SPSS 22.0 统计分析应用教程》，清华大学出版社；参见数据文件：data12-5.sav。）

案例分析：该量表包含 6 个子量表，可分别对 6 个子量表求其 Cronbach 系数和折半系数，之后再对整个量表的信度进行分析。

表 12.21 Oxford 所含子量表情况

子量表	学习策略	题号
子量表 1	记忆策略	1-9
子量表 2	认知策略	10-23
子量表 3	补偿策略	24-29
子量表 4	元认知策略	30-38
子量表 5	情感策略	39-44
子量表 6	社交策略	45-50

12.5 思考与练习

1. 什么是信度和效度？
2. 信度分析包括哪几种，其区别和联系是什么？
3. 某调查表有 $x_1 \sim x_{10}$ 共 10 个变量，经对 14 个对象的调查，其数据如表 12.22 所示。试对此量表进行信度分析。（参见数据文件：data12-6.sav。）
4. 有 20 位选手参加某次书法比赛，主办方邀请了 3 位评委对选手的作品进行评分，评分数据如表 12.23 所示。试分析 3 位评委评分的一致性。（数据来源：冯岩松，《SPSS 22.0 统计分析应用教程》，清华大学出版社；参见数据文件：data12-7.sav。）

表 12.22 某量表的调查数据

序号	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀
1	4	4	4	2	4	4	4	4	5	5
2	5	2	5	3	4	4	3	3	5	5
3	5	5	5	1	5	5	4	2	5	5
4	5	5	4	3	1	2	3	3	4	4
5	5	5	5	3	5	5	4	5	5	5
6	4	4	3	2	3	2	3	4	2	3
7	3	2	1	1	2	1	3	2	4	4
8	3	3	4	2	2	3	1	3	3	3
9	5	5	5	2	3	3	4	5	5	5
10	5	5	5	2	3	3	4	5	5	5
11	4	2	4	3	3	3	3	4	5	4
12	5	5	5	3	3	3	4	4	5	5
13	3	3	5	3	3	5	3	3	5	3
14	5	5	5	1	4	2	4	4	5	5

表 12.23 3 位评委对书法参赛者的评分

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
评委 1	6	4	7	8	2	7	9	7	2	4	6	8	4	3	6	9	9	4	4	5
评委 2	8	5	4	7	3	4	9	8	5	3	9	5	2	3	8	10	8	6	3	3
评委 3	5	6	3	5	3	6	8	5	7	2	5	7	4	6	3	8	6	7	4	6

第 13 章 图表的创建与编辑

大量的统计数据显得纷繁复杂，研究者很难看出其中所蕴涵的信息，而借助于图表，研究者很容易看出图表所体现的数据的分布规律、发展趋势、数量多少和相互关系等信息。图表中包含的信息极多，因为大量数据都能概括在图中，并且一眼就能被理解。俗话说：“一幅图胜过一千个字”，作图有两个主要目的：帮助研究者从数据中提取信息，帮助把信息传给其他人。

SPSS 制图功能很强，可以绘制许多种统计图形，包括条形图、线图、饼图、箱图、直方图以及 3D 图等，这些图形可以由各种统计分析过程生成，也可以直接由图形菜单中所包含的一系列图形过程直接生成。本章主要介绍根据数据直接绘制统计图的过程，即图形菜单中各菜单项所具有的功能。

13.1 SPSS 的图形功能概述

13.1.1 SPSS 创建图形的一般过程

SPSS 图形菜单制作图形可分为 3 个过程：

- (1) 建立数据文件，在数据窗口录入数据，或从其他数据文件中读取数据。
- (2) 利用 SPSS 的图形模块或其他过程生成图形。
- (3) 编辑修饰生成的图形，新生成的图形往往不符合统计图要求，例如图形题目、标尺的单位等，可对其做些调整修饰。

13.1.2 图形生成与数据文件结构

统计图形的生成与数据文件的结构和类型紧密相关。同一数据来源，整理成不同结构的数据文件，在生成图形时，条形图生成对话框中参数设置会有所不同。例如，图 13-1 中是 1~5 月 4 种产品的销售量数据文件，根据此种结构的数据文件要绘制如图 13-2 所示的每种产品在 1~5 月的销售条形图，在进行条形图生成时，参数设置须如图 13-3 所示，即选择图表中的数据为“单独变量的摘要”。

图 13-1 所示的数据也可以整理为图 13-4 所示的数据文件 B，如果要根据此数据文件绘制成与图 13-2 一样的图形，在进行条形图生成时，参数设置须如图 13-5 所示，即选择图表中数据为“个案组摘要”。由此可见，数据文件的整理会直接影响图形生成的参数选择，当不能满足图形生成的要求时，还须对数据文件的结构做适当的调整。

月份	产品1	产品2	产品3	产品4
1	720.00	560.00	411.00	884.00
2	297.00	333.00	860.00	385.00
3	134.00	338.00	464.00	191.00
4	479.00	120.00	849.00	106.00
5	517.00	975.00	847.00	458.00

图 13-1 数据文件 A

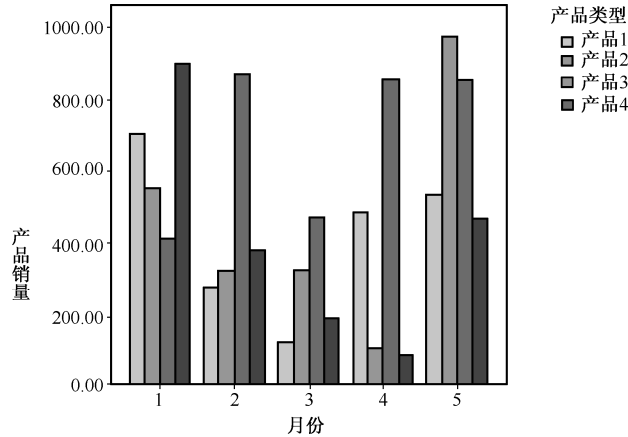


图 13-2 复式条形图



图 13-3 针对数据文件 A 条形图生成参数设置

月份	产品类型	产品销量
1	产品1	720.00
2	产品1	297.00
3	产品1	134.00
4	产品1	479.00
5	产品1	517.00
1	产品2	560.00
2	产品2	333.00

图 13-4 数据文件 B



图 13-5 针对数据文件 B 条形图生成参数设置

13.1.3 图形生成与数据的度量尺度

在第 2 章中介绍了统计数据的标度尺度，由低级向高级分为：名义尺度（Nominal）、定序尺度（Ordinal）、间隔尺度（Scale）。其中，名义尺度的测量水平最低，变量的取值仅代表一定的分类或标识，测量值之间没有大小可言，对应的变量类型可以是数值型，也可以是字符型；定序尺度的测量水平次之，保存测量值之间的一种有序关系，测量值之间有大小关系，但不能做加减等运算，对应的变量类型可以是数值型也可以是字符型；测量水平最高的是间隔尺度，测量值之间可做运算，对应的变量类型只能是数值型。

在定义变量时，须在度量标准栏中设置其度量尺度（也称为测量级别），其在数据分析中的作用不太明显，但在用图表生成器绘制图形时却很重要，度量尺度设置不正确将会影响图形的绘制。例如，在绘制复式条形图时须选择分类变量，分类变量的测量级别只能是名义尺度或定序尺度，对应的变量是数值型或字符型，用数字代表分类（如，0 代表男性，1 代表女性），当变量不是这两种测量级别之一时，将不能作为分类变量使用。

13.2 图表构建器创建图形












13.2.1 图表构建器概述

使用图表构建器创建图形,使用预览模式将图库图表或基本元素拖放到画布(“图表构建器”对话框上“变量”列表右侧的较大区域)上来生成图表,让用户所见即所得,可以提高创建图形的效率,减少一些不可预见的错误。生成图表时,画布会显示图表的预览。虽然预览使用已定义的变量标签和测量级别,但预览并不会显示实际的数据,它使用随机生成的数据简略勾勒出图表的外观。

图表构建器区分不同的测量级别,并根据测量级别的不同以不同方式处理变量。因此,用图表构建器创建图形时,需要正确设置数据文件中各变量的测量级别。通常,在图形生成时将变量分为分类变量和刻度变量,分类变量定义图表中的类别,常用于绘制单独的图形元素或将图形元素分组,刻度变量通常是分类变量的类别汇总。分类变量可以是字符串(字母数值)变量或使用数值代码表示类别的数值变量(例如,0 = Male, 1 = Female),这种数据也称为定性数据,分类变量的测量尺度既可以是名义尺度(Nominal),也可以是定序尺度(Ordinal),刻度变量的测量尺度为间隔尺度(Scale)。

在使用图表构建器创建图形时,变量名称旁会显示该变量测量级别的图标,帮助用户选择该变量作为分类变量,还是作为刻度变量。变量测量级别的图标含义如表 13.1 所示。创建图表时,变量的测量级别很重要,可以在图表构建器中临时更改测量级别,方法是右键单击“变量”列表中的变量,然后选择选项,也可以在数据编辑器的“变量视图”中永久更改变量的测量级别

表 13.1 变量测量级别图标

	测量尺度	数值(N)	字符串(S)	日期	时间
分类变量	间隔尺度 (Scale)		n/a		
	定序尺度 (Ordinal)				
刻度变量	名义尺度 (Nominal)				

13.2.2 使用图表构建器创建图形举例

为介绍使用图表构建程序创建图表操作方法,我们以创建条形图为例来演示,其他类型图表创建的操作方法类似,只需按照图表的统计学意义设置相应选项即可。

【例 13-1】表 13.2 是国民经济与社会发展总量指标中第一、二、三产业在几年中的产值,试绘制条形图对比几年中国国民经济与社会发展总量指标中各产业产值发展趋势及比重。(数据来源:中国统计摘要,2008;参见数据文件: data13-1.sav.)

表 13.2 国民经济与社会发展总量指标 (单位: 亿元)

指 标	1978 年	1990 年	2000 年	2006 年	2007 年
第一产业	1027.5	5062.0	14944.7	24040.0	28095.0
第二产业	1745.2	7717.4	45555.9	103162.0	121381.3
第三产业	872.5	5888.4	38714.0	84721.4	100053.5

创建条形图的具体步骤如下。

第1步 数据组织。

建立数据文件，定义3个变量名：“指标”、“年份”、“指标值”，整理后的数据文件部分截图如图13-6所示，保存为数据文件 data13-1.sav。

第2步 打开“图表构建器”对话框。

选择菜单：“图形→图表构建器”，弹出如图13-7所示的“图表构建器”对话框，该对话框由以下7个部分组成。

指标	年份	指标值
第一产业	1978	1027.5
第二产业	1978	1745.2
第三产业	1978	872.5
第一产业	1990	5062.0
第二产业	1990	7717.4
第三产业	1990	5888.4
第一产业	2000	14944.7

图13-6 整理后的数据文件



图13-7 “图表构建器”对话框

- ① 候选变量框：即左侧变量列表框，如果所选的变量为分类变量，则其下面的“类别”列表会显示该变量的已定义类别。右键单击候选变量框中的某一变量，可以临时更改变量的测量级别、排序规则，显示变量名称或标签名称。
- ② 画布：画布在“图表构建器”对话框的右边区域较大的部分，是生成和预览图表的区域。需要注意的是，画布里的图表不是数据视图里的数据，而是随机产生的数据，简略勾勒图形的外观。
- ③ 图库：即“图库”选项卡，里面预定义了各种常见类型图表，或用户收藏的图表，是常用、高效的作图选择。图库里有条图、线图、饼图/极坐标图、直方图、高-低图、箱图、双轴图等类别，每一类别又包含了多种图表，通过双击或拖放操作，可将图表放置在画布中，供用户进一步添加轴变量或分类变量。
- ④ 基本元素：当图库选项卡提供的图表不能满足用户的特殊需求时，“基本元素”选项卡

提供了从最基本的图表元素作图的素材,如图 13-8 所示,包括一维、二维、三维坐标轴,极坐标轴,双 Y 坐标轴等各种轴,以及点、条、线、区、箱图、高-低图等元素。

⑤ 组/点 ID: 如图 13-9 所示,该选项卡对变量进行聚类、分组设置、行/列面板设置及 ID 标签指定等。行/列面板变量设置就是在行/列上展示多个图表,以便进行对比。



图 13-8 “基本元素”选项卡

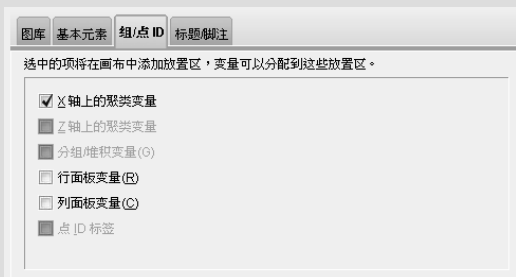


图 13-9 “组/点 ID”选项卡

⑥ 标题/脚注: 如图 13-10 所示,该选项卡对图表进行各级标题、子标题和脚注设置。

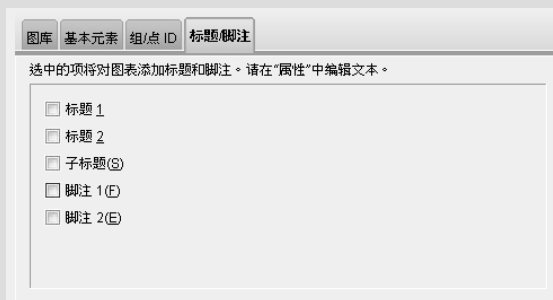


图 13-10 标题/脚注选项卡

⑦ 元素属性: 单击图 13-7 上的“元素属性”按钮打开“元素属性”对话框,如图 13-11 所示。当最初从图库或基本元素将图表放入画布时,“元素属性”对话框也会自动打开。该对话框对图表元素(轴、条、线等)的属性进行设置,如统计量、标签、排序、样式等。

第 3 步 选择图库。

选择“图库”选项卡,双击“条(B)”类别中的第二项“群集条形图”图标,或者直接将“群集条形图”图标拖到画布区域。图库中的每一种图表都是预定义的、由基本元素组成的集合体,因此通过“基本元素”选项卡同样可以创建各种图表,而且比图库更灵活,但通过图库创建图表更快,更简便。

第 4 步 设置图表变量。

尽管画布上有图表,但图表并不完整,因为没有变量或统计量来控制条的高度,以及指定每个条对应的变量类别。图表不能没有变量和统计量。

本例要分析的是几年的产业值比较,所以应按“年份”分类,在“变量”列表框中选择“年份”,将其拖到画布中“是否为 X 轴?”虚线框中作为条形图的 X 轴,并作为分类变量。

因为要比较产业值,所以条形图的条高(即 Y 轴)就是产业的指标值,所以将“指标值”拖放到“计数”蓝色虚线框中(当 X 轴确定后, Y 轴默认统计量为“计数”,当拖放入具体变量时,系统会根据变量的度量尺度自动设置一个统计量)。

由于每一年中又有不同指标的统计值,所以将“指标”变量作为复合分类变量,即在“年

份”分类基础上再做分类,将“指标”拖放到画布右上角的“X轴上的分群:设置颜色”虚线框中。复合分类元素还可以通过“组/点 ID”选项卡添加或取消。

第 5 步 设置元素属性。

图表及变量设置好后,各图表元素属性都是系统默认值,还要根据图表实际需求来修改元素属性,比如统计量、图形样式、排序方式、刻度类型等。

如果“元素属性”对话框没打开,则单击“元素属性”按钮打开,如图 13-11 所示。选择需要编辑的元素,根据不同类别的元素,对话框下面显示不同的属性值或选项。根据需要,修改相应的属性,也可以单击元素列表右边红色的“×”删除元素。

本例元素属性全部采用默认值,不做修改。

第 6 步 设置标题。

通过前面的步骤,图表基本设置完毕,有了坐标及坐标标题、分类标题等,但为了帮助用户理解图表,增强图表的描述语言功能,还可以增加图表的标题和脚注。选择图表创建窗口中的“标题/脚注”选项卡,根据需要选中相应的标题或脚注,然后在“元素属性”对话框选择相应的元素,设置文本内容,单击“应用”按钮。

本例选中标题 1,设置文本内容为“第一、二、三产业各年产值比较图”。

第 7 步 运行结果及说明。

单击“图表构建程序”对话框的“确定”按钮,完成条形图的绘制,得到如图 13-12 所示的结果图形。

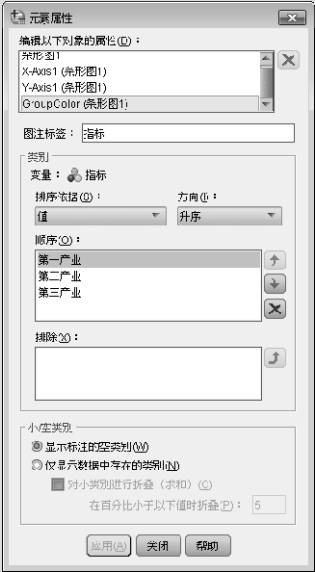


图 13-11 “元素属性”对话框

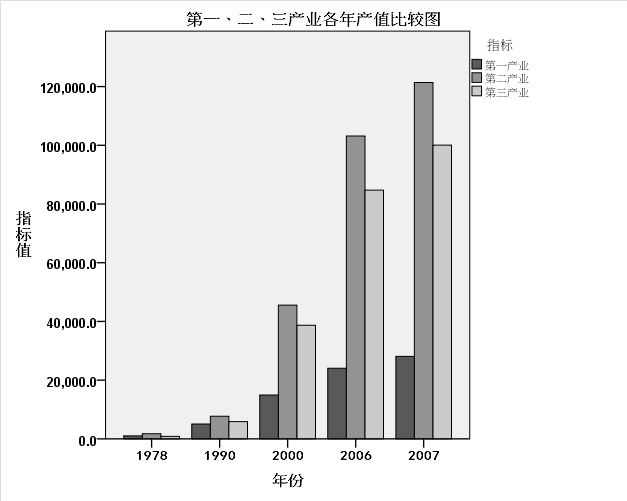


图 13-12 例 13-1 结果图形

图中直条分别表示各年中第一、二、三产业的产值,从图中我们可以得到两个信息:

- (1) 自 1978 年以来,这三种产业的产值都在增加;
- (2) 每年第二产业的产值都最高,第三产业次之,第一产业的产值最少。

13.3 图形画板模板选择器创建图形

13.3.1 图形画板模板选择器概述

用“图表构建器”创建图形时，首先得根据欲创建的图形和数据文件，在图库中选择该图形，再进行轴系的添加，如果用户在创建图形前不清楚应该创建哪种图形来可视化数据，则在图库中选择图形时不知道该选哪种图形，往往不易操作，而“图形画板模板选择器”则与“图表构建器”过程相反，用户可以先选择变量，SPSS 根据变量的类型与个数会自动筛选出可以绘制的图形，用户可以在图形中进行选择。

13.3.2 使用图形画板模板选择器创建图形举例

图形画板模板选择程序是 SPSS 低版本称为“交互式（Interactive）”图表的升级，该程序能创建图表构建程序所能创建的绝大多数统计图表，但与一般的统计图相比，本方式更注重可视化的视觉效果，比如 3D、色彩、动画以及多图对比等。

同样以例 13-1 为例，介绍图形画板模板选择程序创建图表的过程及方法。

第 1 步 准备数据。

打开数据集文件 data13-1.sav。

第 2 步 打开“图形画板模板选择器”对话框。

选择菜单：“图形→图形画板模板选择器”，如图 13-13 所示。该对话框由 4 个选项卡组成，下面分别介绍。



图 13-13 “图形画板模板选择器”对话框

(1) “基本”选项卡：如图 13-13 所示，该选项卡左边是候选变量列表框，列出了数据集中的所有变量，供用户选择一个或多个变量（按住 Ctrl 键）进行可视化表示。SPSS 根据选择变量

的测量级别及其组合，在选项卡的右边列出一些推荐图形，同时，在下面的“摘要”下拉列表中列出可用的统计量。

(2) “详细”选项卡：如图 13-14 所示，当用户设置好“基本”选项卡后，可进一步利用“详细”选项卡对图表的可视化特征进行设置。该选项卡分成以下 3 个区域。

- 可视化类型：通过左上方的下拉列表可对可视化类型重新选择，右边针对选定的可视化类型列出图表的参数（如 X 轴、Y 轴表示的变量）及统计量。
- 可选审美原则：根据选择的图表类型，这里会出现色彩、形状、大小、透明度、资料标记等下拉列表，供用户设置分类变量，系统根据分类变量的各个值，用不同的颜色、形状、大小、透明度来绘制图形元素（如点、条、线等），用分类变量的值来标记图形元素（如散点图的每一个点的值）。
- 面板与动画：面板包括“面板横跨”和“面板向下”，设置一个分类变量，系统根据变量的各个值，在横向/纵向绘制多个子图，供用户进行对比。动画则是系统根据设置的分类变量的各个值，绘制多个子图，在结果浏览窗口中只显示第一个子图，在图形画板可视化编辑窗口的探索模式下，可以像动画一样播放每一个子图，也可以用鼠标拖动播放控制条，逐个子图查看。

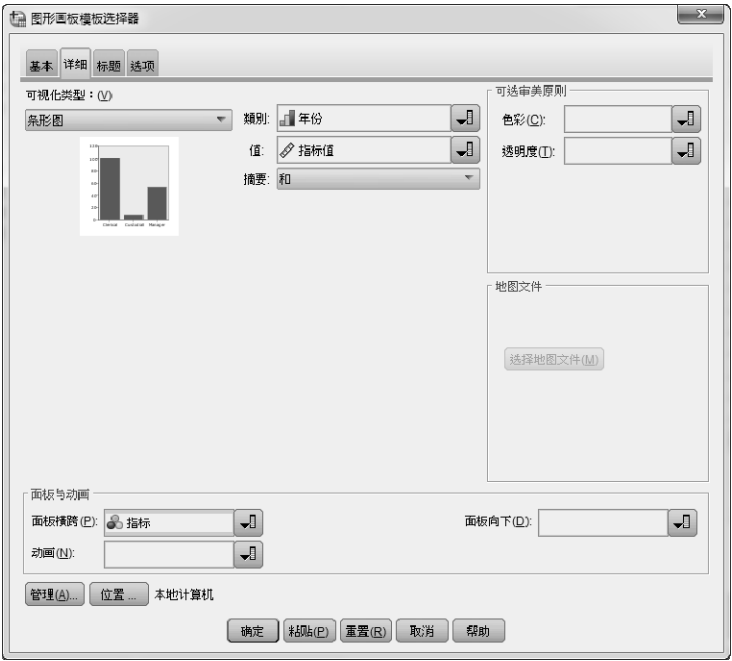


图 13-14 “图形画板模板选择器：详细”选项卡

(3) “标题”选项卡：设置可视化图表的标题，可以使用系统默认标题，也可以用户自定义标题。

(4) “选项”选项卡：设置一些杂项，包括图表在输出窗口的标签、图表使用的系统样式以及对用户缺失值的处理方式等。

第 3 步 设置参数及选项。

按图 13-14 进行设置：选择条形图，类别（即 X 轴）设置为“年份”，值（即 Y 轴）设置为“指标值”，摘要默认为“和”，面板横跨设置为“指标”。

第 4 步 查看运行结果。

单击“确定”按钮，运行结果如图 13-15 所示。输出的图形由 3 个条形图子图组成，可以很好地对比分析第一、二、三产业各年的发展状况。从图中可以看出，第二产业增长速度最快，产值也最高；第一产业增长速度最慢，产值也最低。

类似地，其他设置不变，在面板与动画中，只将面板向下设置“指标”变量，则产生纵向的 3 个条形图子图进行对比；只将动画设置为“指标”变量，则会在结果窗口中默认输出第一个条形图，然后进入可视化编辑窗口，选择探索模式，可以动画显示 3 个条形图。

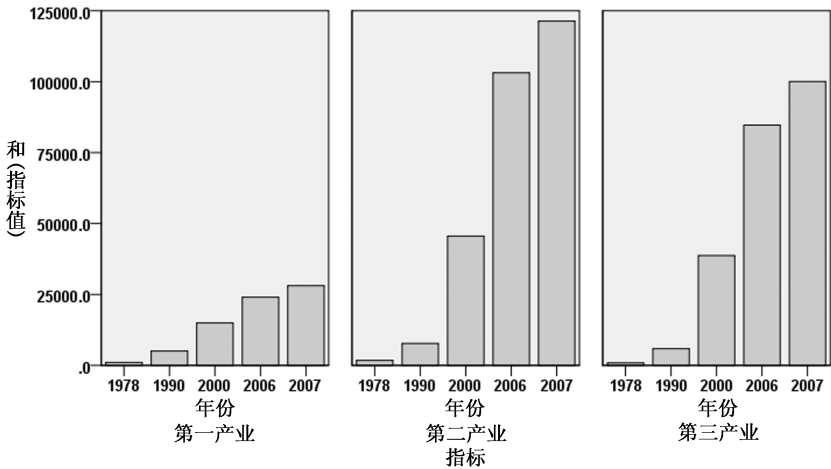


图 13-15 可视化图形输出结果

13.4 使用旧对话框创建图形

13.4.1 条形图

1. 条形图的功能

条形图 (Bar Charts) 描述定类或定序变量的分布，用宽度相等直条的高度来表示非连续性资料的数据大小，用于性质相似的数据进行比较。

2. 条形图的类型

选择“图形→旧对话框→条形图”，弹出如图 13-16 所示的“条形图”对话框，该对话框提供条形图类型的选项，具体而言，包括两个方面的选择，一是条形图类型的选择，二是条形图数据模式的选择。

(1) 类型选择

图 13-16 的“条形图”对话框中提供了 3 种条形图的类型：简单条形图 (Simple)、簇状条形图 (Clustered) 和堆积条形图 (Stacked)。其中简单条形图为默认选项，选中该项则例图外有一个黑框。单击其他例图，可以选择对应的类型。

- 简单条形图 (Simple)，这种类型用有间隔的等宽直条表示各类统计数据的大小，通过它，可以很明显地给出基于某一种分类的各类数据间的对比情况。该类图的形成由两个统计量决定。

- 簇状条形图 (Clustered), 这种类型相当于对简单条形图中的每一个直条对应的数据基于其他变量做进一步的分类, 并且用没有间距的直条表示这些次一级的分类数据, 直条的长度由次一级分类数据所对应的另一个变量的数据大小决定。该类图的形成由 3 个变量决定。
- 堆积条形图 (Stacked), 这种类型也是对简单条形图的一种复合。对于简单条形图中的每一个直条所对应的数据基于某一个变量做进一步的分类, 然后用进行次一级分类之后各类数据的大小占总直条对应数据的大小的比例关系将原直条划分为多个段, 并用不同的颜色或阴影填充方式来表示这种分段。这样形成的图在形式上就好像堆垒条形积木一样, 因此称其为堆积条形图。

(2) 条形图模式的选择

在图 13-16 的“图表中的数据为”选项框中提供了条形图数据模式的选项, 具体含义如下:

- 个案组摘要。此选项为默认选项, 表示统计量按个案分组方式组织, 即将根据分组变量对所有个案进行分组, 然后根据分组后的个案数据创建条形图。
- 单独变量的摘要。以变量的数据作为分组模式, 表示将根据每个变量的数据创建条形图。
- 单个个案的值。个案模式, 表示将为分组变量中的每一个个案生成一个条形图, 条带的长度表示观测值的大小。当数据文件中包含大量个案时, 不适宜用个案模式条形图来描述。设置好以上两种选项以后, 单击“确定”按钮, 可以做相关图的进一步设置。



图 13-16 “条形图”对话框

3. 条形图的生成

同样以例 13-1 为例, 介绍条形图的创建过程及方法。

第 1 步 准备数据。

打开数据集文件 data13-1.sav, 部分数据如图 13-6 所示。

第 2 步 打开“条形图”对话框。

选择菜单: “图形→旧对话框→条形图”, 弹出如图 13-16 所示的“条形图”对话框, 图形类型选择“簇状”, 因为要生成各个年份每年第一、二、三产业产值的对比条形图, 根据数据的组织形式, 条形图数据模式选择“个案组摘要”, 单击“定义”按钮弹出如图 13-17 所示的“定义簇状条形图: 个案组摘要”的对话框。

第 3 步 确定分类变量。

要分析的是几年的产业值比较, 则应按年份分类, 在对话框左侧的列表框中选择“年份”变量名, 单击向右箭头按钮, 将其添加到“类别轴”框中。

第 4 步 选择分类变量中聚类定义依据。

聚类定义依据变量的选择, 是在第 3 步确定的分类基础上再做分类的变量, 每一年中又有不同指标的统计值, 故将“指标”变量作为复合分类变量, 将其添加到“聚类定义依据”框中。

第 5 步 确定直条表示的方式和统计量。

在“条形表示”选项组中, 给出确定条形图中直条的长度代表的统计量, 用“指标值”作为直条代表的值, 选择“其他统计”单选项, 激活下面的变量框, 将“指标值”移入其中, 系统自动计算平均值作为直条的长度。如果不希望对变量值取平均值, 可以单击下面的“更改统计...”按钮则可进行修改, 此处不再赘述。



图 13-17 “定义簇状条形图：个案组摘要”对话框

第 6 步 选择条形图的标题和脚注。

如果需要的话，可以单击图 13-17 中的“标题”按钮，弹出标题对话框，输入条形图的标题和脚注，在“标题”选项组的“第一行标题”文本框中输入本例的标题“第一、二、三产业在每年的平均产值比较图”。

第 7 步 运行结果及说明。

单击图 13-17 的“确定”按钮，完成条形图的绘制，得到与图 13-12 相同的结果图形。

13.4.2 三维条形图

1. 三维条形图的功能

三维条形图（3-D Bar）允许用户绘制具有两个分类轴的条形图。例如，不同性别和学历的人员的收入水平可以表示在一个三维条形图中，其中性别和学历可以作为两分类变量，而相应人群的平均收入水平作为条形图直条的大小。

2. 三维条形图的类型

选择“图形→旧对话框→三维条形图”，弹出如图 13-18 所示的“三维条形图”对话框，该对话框提供了两组选项：“X 轴表示”选项组中给出横轴模式选择的 3 个选项，“Z 轴表示”选项组给出相同的选项。

分别在两选项组中选择相应模式即可组成三维条形图的类型，虽然两选项组中各提供 3 个选

项，可构成 9 种不同的选择对，但只有下面的 4 种选择对才是 SPSS 允许的配对类型，即三维条形图的类型，如表 13.3 所示。

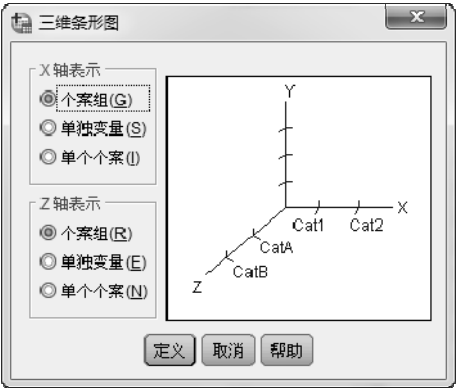


图 13-18 “三维条形图”对话框

表 13.3 三维条形图的类型

X 轴的表示	Z 轴的表示	三维条形图类型
个案组	个案组	个案分组模式
个案组	单独变量	组间变量分类模式
个案组	每个个案	组内个案模式
单独变量	每个个案	变量个案模式

图 13-19 给出 4 种类型的示意图，其中 A 表示个案分组模式 (Summaries for Groups of Cases)，B 表示组间变量分类模式 (Summaries of Separate Variables by Group)，C 表示组内个案模式 (Values of Individual Cases in Groups)，D 表示变量个案模式 (Values of Individual Cases for Separate Variables)。

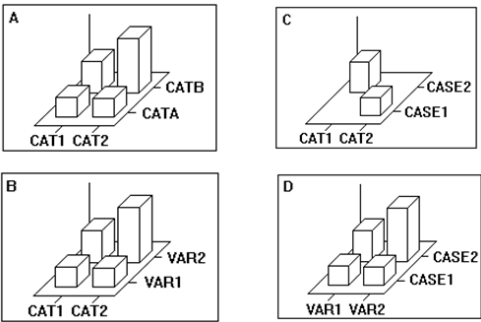


图 13-19 4 种三维条形图类型示意图

3. 三维条形图的生成

【例 13-2】 表 13.4 是每百户城乡居民家庭在 2005 年和 2006 年拥有耐用消费品的数量，试绘制各年城乡居民拥有不同耐用消费品的三维条形图。（数据来源：成都统计年鉴，2007；数据文件：data13-2.sav。）

表 13.4 每百户城乡居民家庭年末耐用消费品拥有量

耐用消费品	地区	单位	2005 年	2006 年
空调机	城市	台	103.8	114.3
	农村	台	6.2	7.6
洗衣机	城市	台	98.3	99
	农村	台	74.5	79.5
电冰箱	城市	台	96.5	98
	农村	台	31.8	36.1
抽油烟机	城市	台	54.5	57
	农村	台	4.9	4.7
淋浴热水器	城市	台	96.5	97.5
	农村	台	26.8	29.2
彩色电视机	城市	台	145.8	151.25
	农村	台	107.6	113

根据表 13.4 的统计数据建立数据文件，定义 4 个变量：“消费品”、“地区”、“年份”、“数量”，整理后的数据文件如图 13-20 所示。

绘制三维条形图的具体步骤如下。

第 1 步 数据组织。

打开如图 13-20 所示的数据文件 data13-2.sav。

第 2 步 打开主对话框。

选择“图形→旧对话框→三维条形图”，弹出如图 13-18 所示的“三维条形图”对话框，“X 轴表示”选项组中选择“个案组”，“Z 轴表示”选项组中选择“个案组”，单击“定义”按钮弹出如图 13-21 所示的主对话框。

消费品	地区	年份	数量
空调机	城市	2005	103.8
空调机	农村	2005	6.2
空调机	城市	2006	114.3
空调机	农村	2006	7.6
洗衣机	城市	2005	98.3
洗衣机	农村	2005	74.5
洗衣机	城市	2006	99.0
洗衣机	农村	2006	79.5
电冰箱	城市	2005	96.5

图 13-20 例 13-2 数据文件部分数据



图 13-21 “定义三维条形图：个案组摘要”对话框

第 3 步 确定分类变量。

将消费品的种类和年份分别作为分类变量,将变量“年份”移入“X 类别轴”框,变量“耐用消费品”移入“Z 类别轴”框。

第 4 步 选择分类的聚类变量。

聚类变量是在第 3 步确定的分类中再做分类的变量,在三维条形图中,X 分类轴和 Z 分类轴都可以设置聚类变量,本例中将“地区”变量作为 X 分类轴的聚类变量,选择该变量添加到“X 中的聚类”框中。

第 5 步 确定条形表示的方式和统计量。

在“条形表示”下的组合框中选择“值的总和”,激活下方的“变量”框,将“数量”添加到“变量”框中。

第 6 步 添加标题并完成图形绘制。

若要添加标题则单击“标题...”按钮进行设置,单击“确定”按钮完成三维条形图的绘制,按如上步骤绘制的图形如图 13-22 所示。

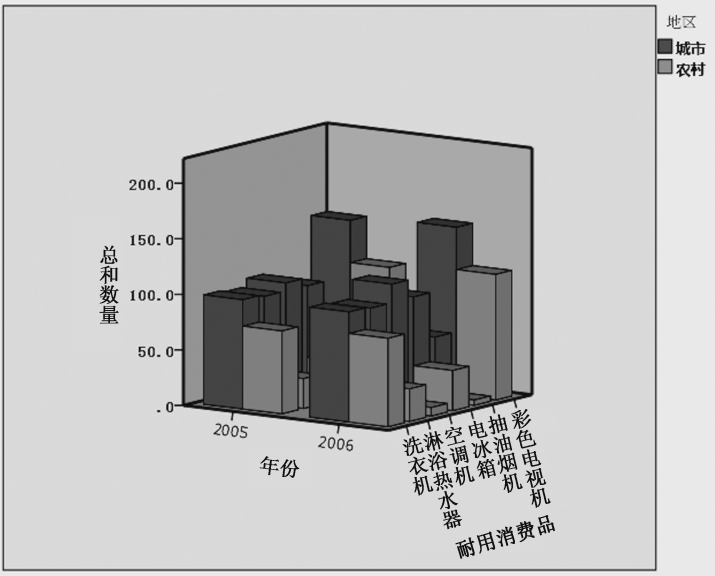


图 13-22 例 13-2 数据绘制的三维条形图

13.4.3 折线图

1. 折线图的功能

折线图 (Line Charts) 是用线条的上下波动的形式,来反映连续性的相对资料的变化趋势。它主要用于表示现象在时间上的变化趋势、现象的分配情况和两个现象之间的依存关系等。

2. 折线图的类型

选择“图形→旧对话框→折线图”,弹出如图 13-23 所示的“折线图”对话框,在该对话框中提供了定义线图的 3 种类型。

- 简单折线图 (Simple Line Chart): 用一条折线表示某个现象的变化趋势;
- 多线折线图 (Multiple Line Chart): 用多条折线同时表示多种现象的变动趋势;

➤ 垂直折线图（Vertical Line Chart）：反映某些现象在同一时期内差距的统计图。

与条形图类似，除了选择线图的类型外，还要选择对话框的“图表中的数据为”选项组中的折线图的模式，具体含义与条形图的 3 种分类相同。



图 13-23 “折线图”对话框

3. 折线图的生成

【例 13-3】 表 13.5 是在几年间统计的邮电业务基本情况的部分数据，试绘制每种业务在这几年中发展情况的折线图，以比较每种业务在这几年中的变化。（数据来源：中国统计摘要 2008；数据文件：data13-3.sav。）

表 13.5 邮电业务基本情况

指 标	单位	1990 年	1995 年	2000 年	2006 年	2007 年
特快专递	万件	343	5563	11031	26988.0	120053.1
移动电话年末用户	万户	1.8	362.9	8453.3	46105.8	54728.6
固定电话年末用户	万户	685.0	4070.6	14482.9	36778.6	36544.8

绘制折线图的具体步骤如下：

第 1 步 数据组织。

根据表 13.5 的统计数据建立数据文件，定义三个变量名：指标分类、年份、指标值，整理后的 SPSS 数据文件如图 13-24 所示。

指标分类	年份	指标值
固定电话年末用户	1990.0	685.0
特快专递	1990.0	343.3
移动电话年末用户	1990.0	1.8
固定电话年末用户	1995.0	4070.6
特快专递	1995.0	5562.7
移动电话年末用户	1995.0	362.9
固定电话年末用户	2000.0	14482.9
特快专递	2000.0	11031.4

图 13-24 数据文件部分数据

第2步 打开定义折线图的对话框。

选择“图形→旧对话框→折线图”，弹出图 13-23 所示的“折线图”对话框，选择“多线”图形，“图表中的数据为”选项组中选择“个案组摘要”选项，单击“定义”按钮弹出如图 13-25 所示的“定义多线折线图：个案组摘要”对话框。

第3步 选择分类变量。

将“年份”作为分类变量，从左边的变量列表中选择“年份”添加到“类别轴”框中。每种邮电业务指标是一条折线，于是从左边的变量列表中选择变量“指标分类”添加到“折线定义依据”框中。

第4步 确定折线表示的方式和统计量。

在“折线表示”选项组中给出确定折线代表的统计量，选择“其他统计（例如平均值）”单选项，激活下面的变量框，将“指标值”移入其中，操作方式与条形图类似。



图 13-25 “定义多线折线图：个案组摘要”对话框

第5步 运行结果及说明。

单击图 13-25 的“确定”按钮，完成多重线图的绘制，得到如图 13-26 所示的结果图形。从图中可以得到如下信息：

- (1) 从 1990 年开始，特快专递、移动电话业务呈逐年上升的趋势，特别是特快专递业务，在 2006 至 2007 年间，业务增长迅猛；
- (2) 固定电话业务在 1990 年至 2006 年间呈增加趋势，但在 2006 年至 2007 年间有下降的趋势。

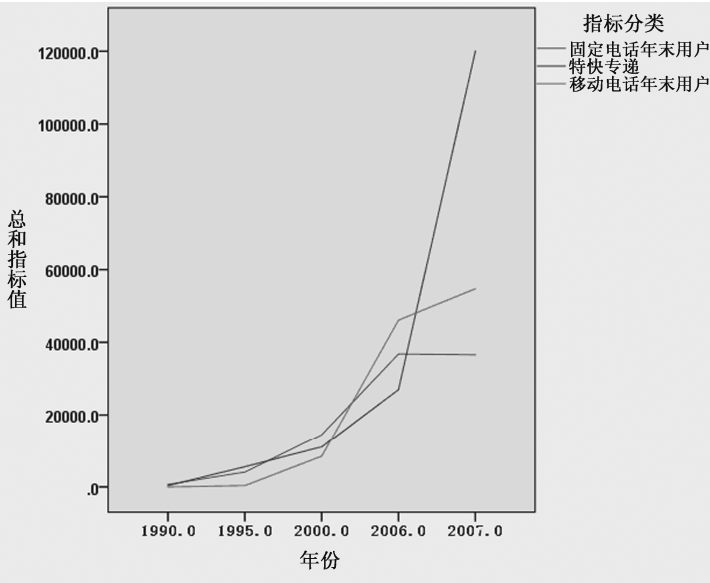


图 13-26 折线图结果图形

13.4.4 面积图

面积图（Area Charts）是用线段下的阴影面积来强调现象变化的统计图。面积图使用面积来表示连续性的频数分布资料，面积越大，频数越多，反之亦然。面积图有两种类型：一种是简单面积图（Simple），是用面积的变化表示某种现象变动趋势的统计图；另一种是堆栈面积图（Stacked），是用不同种类的面积表示多种变动趋势和总体内部构成的统计图。

条形图、折线图和面积图三者都是用来描述变量的分布情况的，并且可以相互转换。面积图的定义与前面二者类似，这里不再详细描述。

图 13-27 的堆栈面积图是用例 13-1 中的数据绘制得到，从图中可以很容易地看出第二产业在 2006 到 2007 年期间产值增长很快。

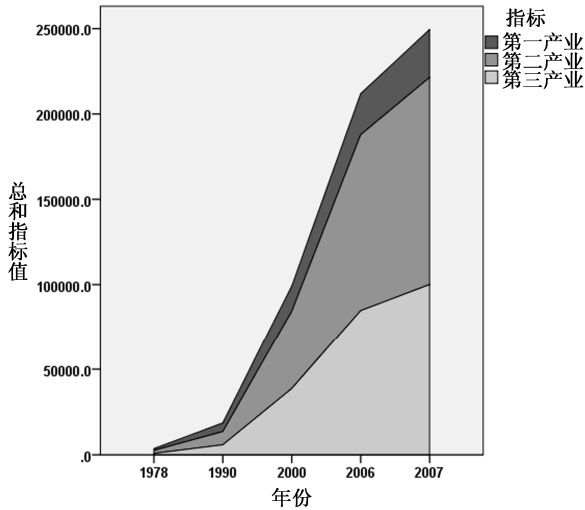


图 13-27 堆栈面积图

13.4.5 饼图

1. 饼图的功能

饼图 (Pie Charts) 也称作圆图, 是用圆的整体面积代表被研究对象的总体, 按各构成部分的比重把圆面积分成若干个扇形, 用以表示对象的部分与总体的比例关系的统计图。

2. 饼图的类型

选择“图形→旧对话框→饼图”, 弹出如图 13-28 所示的“饼图”对话框, 在该对话框中提供了饼图对应的 3 种数据模式:

- 个案组摘要;
- 单独变量摘要;
- 单个个案的值。



图 13-29 显示创建的对应以上三种模式的饼图, A 表示个案分组模式, B 表示变量分组模式, C 表示个案模式。

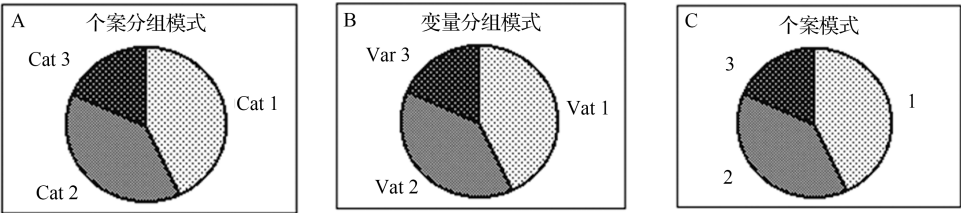


图 13-29 饼图的 3 种模式示意图

3. 饼图的生成

【例 13-4】表 13.6 是西部地区 2007 年末各省份人口数据, 试绘制各省份占西部地区人口比例的饼图。(数据来源: 中国统计摘要, 2008; 数据文件: data13-4.sav。)

表 13.6 各地区年末总人口 (单位: 万人)

地 区	人口
重 庆	2816
四 川	8127
贵 州	3762
云 南	4514
西 藏	284

绘制饼图的具体步骤如下:

第 1 步 数据组织。

根据表 13.6 的统计数据建立数据文件, 定义两个变量: “地区”、“人口”, 整理后的 SPSS 数据文件保存为 data13-4.sav。

第 2 步 打开“定义饼图: 个案组摘要”对话框。

在如图 13-28 所示的对话框中选择“个案组摘要”, 单击“定义”按钮, 弹出如图 13-30 所示的对话框。

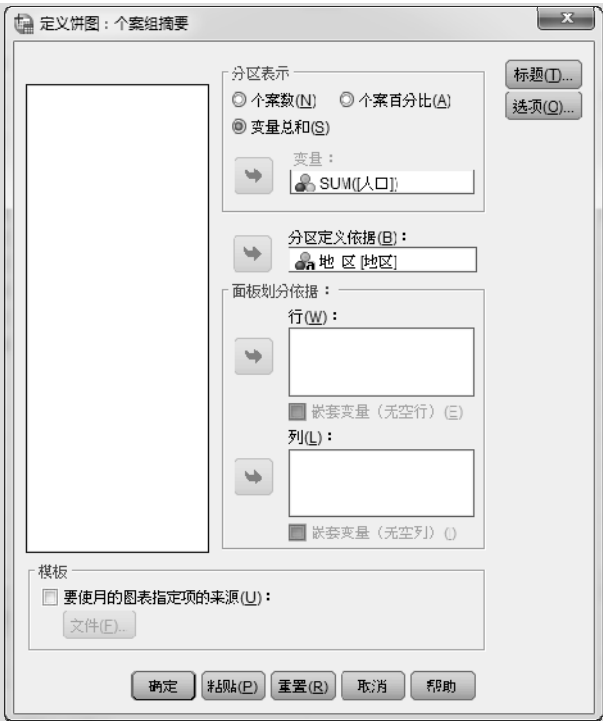


图 13-30 “定义饼图：个案组摘要”对话框

第 3 步 选择分类变量。

从左边的变量列表框中选择变量“地区”作为分类变量，添加到“分区定义依据”框中，在“分区表示”选项组中选择“变量总和”选项，激活“变量”框，将变量“人口”作为分区代表的含义添加到此框中。

第 4 步 运行结果及说明。

单击图 13-30 的“确定”按钮，完成饼图的绘制，得到如图 13-31 所示的结果图形。

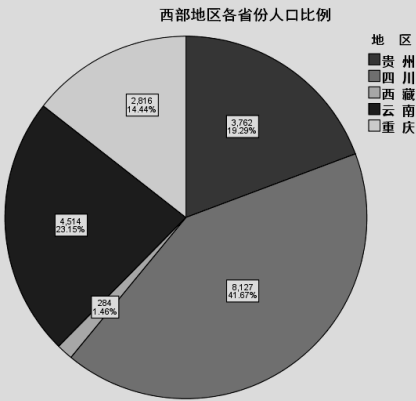


图 13-31 例 13-4 结果图

从图中可以一目了然地看出，西藏地区的人口在西部地区所占比例最小。

13.4.6 盘高-盘低图

1. 高低图的功能

高低图 (High-Low Charts) 是一种说明某种现象在单位时间内变化情况的统计图, 它适合描述每小时、每天、每周等时间内不断波动的市场信息资料。例如股票、商品价格等, 高低图既说明某些现象在短时间内的变化, 也说明它们长期的变化趋势。

2. 高低图的类型

选择 “图形→旧对话框→高低图”, 弹出图 13-32 所示的 “盘高-盘低图” 对话框, 在该对话框中提供了盘高-盘低图对应的 5 种类型。

- 简单盘高-盘低-收盘图: 该图利用小方框表示某段时间内的最终数值, 用小方框上下的触须表示该段时间内取值的最大值和最小值。这种图形适合用于股票、期货和外汇金融等, 它说明每天的最高价格、最低价格和收盘时的价格。
- 简单范围条形图: 这种图形用长条的长度代表每个时间段最高值与最低值之差。
- 簇状盘高-盘低-收盘图: 与简单盘高-盘低-收盘图类似, 但是它可以同时描述两种或两种以上证券或期货的价格情况。
- 簇状范围条形图: 与简单范围条形图类似, 但是可以描述两个或两个以上证券或金融的情况。
- 差别面积图: 这种图形利用不同的曲线表示同一段时间内的两种不同情况, 并且用阴影填充曲线之间的区域。



图 13-32 “盘高-盘低图”对话框

在图 13-32 的 “图表中的数据为” 选项组中提供了定义盘高盘低图的 3 种数据组织模式, 具体含义和条形图的 3 种模式相同: 个案组摘要、单独变量的摘要、单个个案的值。

3. 盘高-盘低图的生成

【例 13-5】表 13.7 是某股票在一段时间的交易数据, 试绘制该股票的简单盘高-盘低图。(数据文件: data13-5.sav)

表 13.7 某股票的交易数据

交易日期	最高价	最低价	收盘价
9 月 1 日	65.17	53.8	64.38
9 月 2 日	57.51	51.66	56.5
9 月 3 日	56.99	51.47	51.85
9 月 4 日	69.01	62.16	63.5
9 月 5 日	66.84	63.82	65.13
9 月 6 日	63.77	60	62.42
9 月 7 日	65.08	55.25	57.31
9 月 8 日	65.05	55.17	54.44

绘制高低收盘图的具体步骤如下。

第 1 步 数据组织。

根据表 13.7 的数据建立数据文件，定义 4 个变量名：交易日期、最高价、最低价、收盘价，保存为数据文件 data13-5.sav。

第 2 步 打开定义高低收盘图的对话框。

在如图 13-32 所示的对话框中选择“简单盘高-盘低-收盘图”，选择“单独变量的摘要”数据组织模式，单击“定义”按钮，弹出如图 13-33 的对话框。



图 13-33 “定义简单盘高-盘低-收盘图：单独变量的摘要”对话框

第 3 步 选择分类变量。

从左边的变量列表框中选择变量“交易日期”作为分类变量，添加到“类别轴”框中。

第 4 步 确定条带高度代表的含义。

设置“条形表示”选项组的各项值，“最高价”添加到“高”框中，“最低价”添加到“低”框中，“收盘价”移入“闭合”框中。

第 5 步 运行结果及说明。

单击图 13-33 的“确定”按钮，完成盘高-盘低-收盘图的绘制，得到如图 13-34 所示的结果图形。

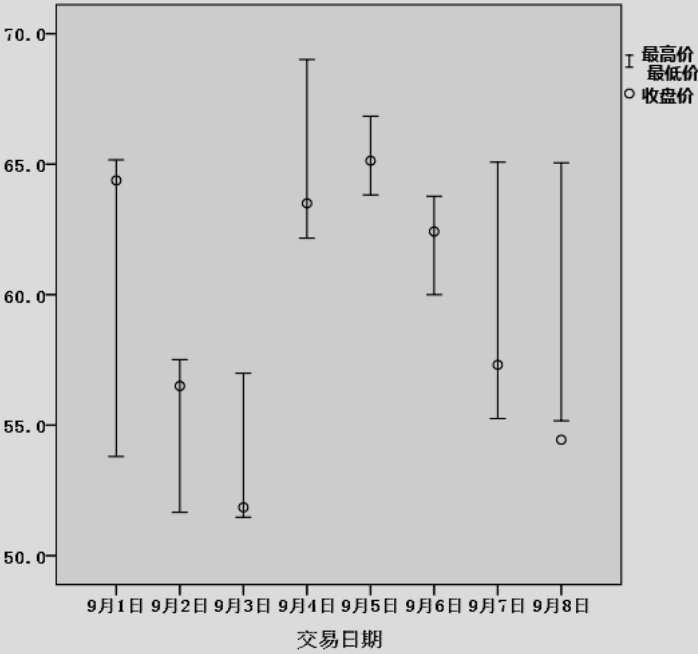


图 13-34 例 13-5 结果图

从图中可以看出在每一个交易日，该股票价格的波动范围和收盘价。

13.4.7 箱图

1. 箱图的功能

箱图 (Boxplot) 又称为箱形图，它是一种用来描述数据分布的统计图形，它可以用来表示观测数据的中位数、4 分位数和极值等描述性统计量，从视觉的角度观测变量值的分布情况。

2. 箱图的类型

选择“图形→旧对话框→箱图”，弹出如图 13-35 所示的“箱图”对话框，在该对话框中提供了箱图对应的两种类型：简单箱图和簇状箱图。

在“图表中的数据为”选项组中提供了定义箱图的两数据组织模式：个案组摘要和单独变量的摘要。



图 13-35 “箱图”对话框

3. 箱图的生成

【例 13-6】表 13.8 是某班各科成绩数据,根据该数据绘制各科成绩的简单箱图。(数据文件: data13-6.sav.)

绘制箱图的具体步骤如下。

第 1 步 数据组织。

根据表 13.8 的数据建立数据文件,定义 3 个变量:“班级”、“科目”、“成绩”,数据文件的部分数据如图 13-36 所示,保存为数据文件 data13-6.sav。

表 13.8 某班成绩表

班级	语文	数学	英语
1	83	73	85
1	74	11	91
1	73	16	11
1	30	75	55
1	60	56	32
1	95	19	56

班级	科目	成绩	班级	科目	成绩
1	语文	83	1	数学	75
1	语文	74	1	数学	56
1	语文	73	1	数学	19
1	语文	30	1	英语	85
1	语文	60	1	英语	91
1	语文	95	1	英语	11
1	数学	73	1	英语	55
1	数学	11	1	英语	32
1	数学	16	1	英语	56

图 13-36 例 13-6 数据文件部分数据

第 2 步 打开定义箱图的对话框。

在如图 13-35 所示的对话框中选择“简单”箱图类型,选择“个案组摘要”数据组织模式,单击“定义”按钮,弹出如图 13-37 所示的对话框。

第 3 步 选择分类变量。

在图 13-37 所示的对话框中,从左边的变量列表框中选择变量“科目”作为分类变量,添加到“类别轴”框中。

第 4 步 确定绘制箱图的变量。

将“成绩”变量移入“变量”框中,如图 13-37 所示。

第 5 步 运行结果及说明。

单击图 13-37 中的“确定”按钮,完成箱图的绘制,得到如图 13-38 所示的结果图形。



图 13-37 “定义简单箱图：个案组摘要”对话框

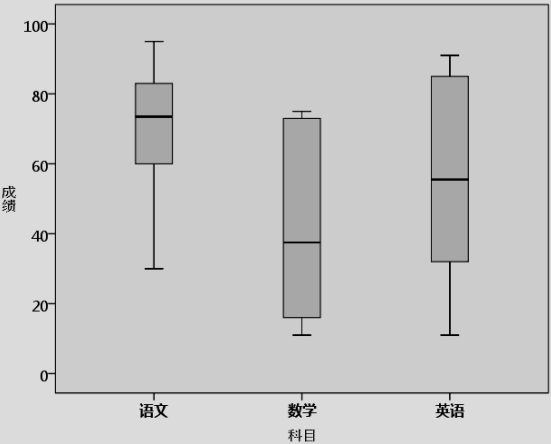


图 13-38 例 13-6 结果图

图 13-38 是关于各科成绩的箱图，图中每个箱形的含义为：从底部开始的线段到矩形框包含了 1/4 的观测数据，从矩形框的低端到矩形框中间的线段包含 1/4 的观测数据，从这条线到矩形框结束又包含 1/4 的观测数据，从矩形框结束到顶端的线段包含 1/4 的观测数据。当最小或最大的观测值距箱形的距离比箱形本身的长度要大好几倍时，箱图中箱形外的线并不一定是从最小的观测值开始并到最大的观测值结束。在这种情况下两端用点标上观测值即可，这种值被称为离群值，也就是非正常值。

13.4.8 误差条图

1. 误差条图的功能

误差条图（Error Bar）是一种描述数据总体离散情况分布的统计图形，可以反映数据的离散情况，并且描绘正态分布资料的描述性指标，如均值、标准差，并由此求得参数数值范围、总体均值的置信区间等。

2. 误差条图的类型

选择“图形→旧对话框→误差条形图”，弹出如图 13-39 所示的“误差条形图”对话框，在该对话框中提供了误差条图对应的两种类型：

- 简单误差条图。对分类轴变量的每个类型生成一个分布误差条形图。
- 簇状误差条图。对分类轴上的变量的每一类型生成一簇误差条形图，每一簇中误差条形图将区分变量显示。



图 13-39 “误差条形图”对话框

在“图表中的数据为”选项组中提供了定义误差条图的两组数据组织模式：个案组摘要模式和单独变量的摘要模式。

3. 误差条图的生成

【例 13-7】 绘制表 13.8 中各科成绩的简单误差条图。（数据文件：data13-6.sav。）

绘制图形的数据文件同图 13-36，绘制误差条形图的具体步骤如下。

第 1 步 打开数据文件 data13-6.sav。

第 2 步 打开定义误差条图的对话框。

在如图 13-39 所示的对话框中选择简单误差条图类型，选择“个案组摘要”数据组织模式，单击“定义”按钮，弹出如图 13-40 所示的对话框。

第 3 步 选择分类变量。

从左边的变量列表框中选择变量“科目”作为分类变量，添加到“类别轴”框中。

第 4 步 确定绘制误差条图的变量。

将“成绩”变量添加到“变量”框中。

第 5 步 确定误差条图中条带的含义。

“条形表示”选项组用于选择误差条图中条带的含义，在下拉列表框中有 3 个选项：

- 平均值的置信区间选项：该选项为系统默认选项，表示以平均值的置信区间表示条带的含义。如果选择此项，则可以在下方的“级别”框中设置置信区间，系统默认为 95%；
- 平均值的标准误差选项：表示以均值的标准误差作为条带的含义；
- 标准差选项：表示以标准差作为条带的含义。

本例采用系统默认选项。

第 6 步 运行结果及说明。

单击图 13-40 的“确定”按钮，完成误差条形图的绘制，得到如图 13-41 所示的结果图形。



图 13-40 “定义简单误差条形图：个案组摘要”对话框

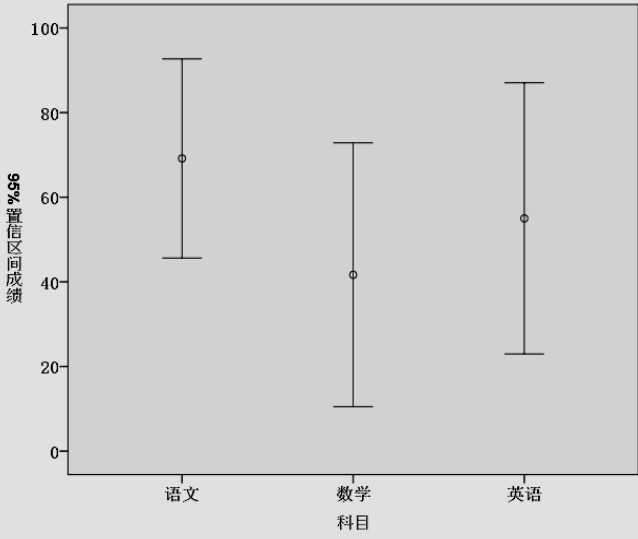


图 13-41 例 13-7 结果图

从图 13-41 中可以观察到各科成绩在置信度为 95% 的成绩置信区间的误差条图。图中的圆点表示平均数，上下两条横线表示置信区间的上下限和标准误差。

13.4.9 人口金字塔图

1. 人口金字塔图的功能

人口金字塔图 (Population Pyramid) 利用图形直观描述分类变量中不同分类的某种属性在各个区间取值的频数。例如, 医疗机构对某种药物进行对比试验, 得到服药和没服药的人员的体重数据, 我们可以将两类人员体重分布的频数分别表示成长条状, 然后依次罗列在一条坐标纵轴上, 形成一塔图。利用金字塔图可以直观地表示出某种属性的人员的变化规律。

2. 金字塔图的生成

【例 13-8】表 13.9 是 2007 年 80 岁以上人口在各年龄段的数据统计, 试绘制不同性别各年龄段的人口数量金字塔图。(数据来源: 中国人口和就业统计年鉴, 2007; 数据文件: data13-7.sav。)

表 13.9 2007 年 80 岁以上各段年龄人口统计

年龄	男 (人数)	女 (人数)	年龄	男 (人数)	女 (人数)
80	1254	1629	88	213	371
81	1034	1393	89	130	286
82	946	1253	90	129	233
83	723	1078	91	76	170
84	647	867	92	62	164
85	535	798	93	55	113
86	376	645	94	44	84
87	307	481	95+	75	204

绘制金字塔图的具体步骤如下。

第 1 步 数据组织。

根据表 13.9 的数据建立 SPSS 数据文件, 定义 3 个变量: “年龄”、“性别”、“人口”, 数据文件如图 13-42 所示, 保存为 data13-7.sav。

年龄	性别	人口	年龄	性别	人口	年龄	性别	人口
86	女	645	91	男	76	80	男	1254
87	女	481	92	男	62	81	男	1034
88	女	371	93	男	55	82	男	946
89	女	286	94	男	44	83	男	723
90	女	233	95+	男	75	84	男	647
91	女	170	80	女	1629	85	男	535
92	女	164	81	女	1393	86	男	376
93	女	113	82	女	1253	87	男	307
94	女	84	83	女	1078	88	男	213
95+	女	204	84	女	867	89	男	130
			85	女	798	90	男	129

图 13-42 例 13-8 数据文件

第 2 步 打开定义金字塔图的对话框。

选择“图形→旧对话框→人口金字塔”, 弹出如图 13-43 所示的“定义人口金字塔”对话框。



图 13-43 “定义人口金字塔”对话框

第 3 步 选择分类变量和属性变量。

从左边的变量列表框中选择“性别”变量作为拆分的分类变量，移入“拆分依据”框中，选择“年龄”变量作为描述分类变量属性的变量，移入“显示基于下列各项的分布”框中。

第 4 步 确定计算频数的方式。

在“计数”选项组中确定计算频数的方式，有两个选项：

- 根据数据计算计数，此选项为默认选项；
- 从变量中获取计数。选择该项，表示计数值保存在某一变量中。

本例中计数值放在“人口”变量中，所以选择第二项，并将“人口”变量移入“变量”框中。

第 5 步 运行结果及说明。

单击图 13-43 的“确定”按钮，完成金字塔图的绘制，得到如图 13-44 所示的结果图形。

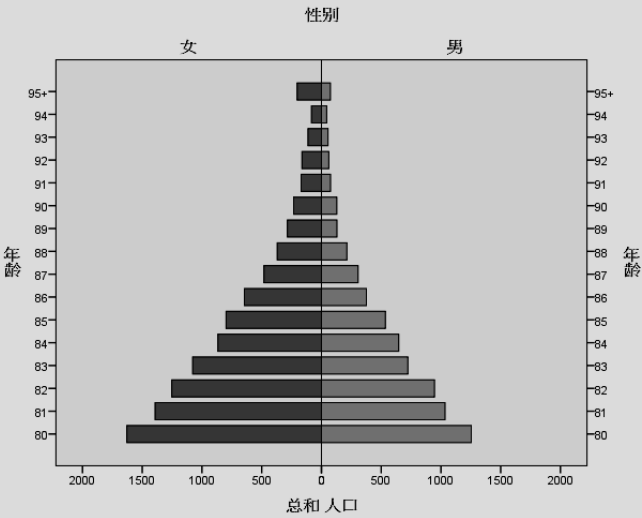


图 13-44 例 13-8 结果图

从图 13-44 中可以观察到：在同一个年龄层，女性的人数多于男性的人数，95 岁以上年龄段特别突出；年龄层越往上，人口数量越少，但 95 岁以上年龄段是个例外。

13.4.10 散点图

1. 散点图的功能

散点图又称散布图或相关图，它是以点的分布反映变量之间相关情况的统计图形，根据图中各点分布走向和密集程度，判定变量之间协变关系的类型。具体创建过程是利用在二维或三维空间中绘制出两个或三个变量确定的点，然后通过这些点的分布特征来显示数据分布特征。

2. 散点图的类型

选择“图形→旧对话框→散点图/点图”，弹出如图 13-45 所示的“散点图/点图”对话框，在该对话框中提供了散点图对应的 5 种类型。

- 简单散点图：描述两个变量之间的关系；
- 重叠散点图：利用将两幅简单散点图叠加到一张图上的形式同时描述多个变量之间的两两关系；
- 矩阵散点图：利用类似矩阵的形式，在一张图上同时描述多个变量之间的两两关系；
- 三维散点图：描述三个变量之间的相互关系；
- 简单点图：描述一个变量在各个值的分布情况。



图 13-45 “散点图/点图”对话框

3. 散点图的生成

【例 13-9】图 13-46 是 5 岁儿童体重、身高、胸围的部分 SPSS 数据，试绘制儿童体重与身高、体重与胸围的重叠散点图。（数据文件：data13-8.sav。）

绘制重叠散点图的具体步骤如下。


第 1 步 数据组织。

数据文件如图 13-46，定义 3 个变量：“体重”、“身高”、“胸围”，保存为 data13-8.sav。

第 2 步 打开定义重叠散点图的对话框。

在如图 13-45 所示的对话框中选择“重叠散点图”类型，单击“确定”按钮弹出图 13-47 所示的“重叠散点图”对话框。

第 3 步 确定重叠散点图的配对变量。

在如图 13-47 所示的对话框中分别选择变量“体重”与“身高”、“体重”与“胸围”，移入“Y-X 对”列表框，作为重叠散点图的两个配对，利用  按钮可以改变变量配对的 Y-X 轴顺序。

体重	身高	胸围
17	110.6	55
15	103.2	50
21	112.5	55
16	106.8	50
18	109.7	56
20	111.1	55
17	105.8	51
17	109.5	53
18	109.2	53

图 13-46 例 13-9 部分数据

第 4 步 运行结果及说明。

单击图 13-47 的“确定”按钮，完成重叠散点图的绘制，得到图 13-48 的结果图。



图 13-47 “重叠散点图”对话框

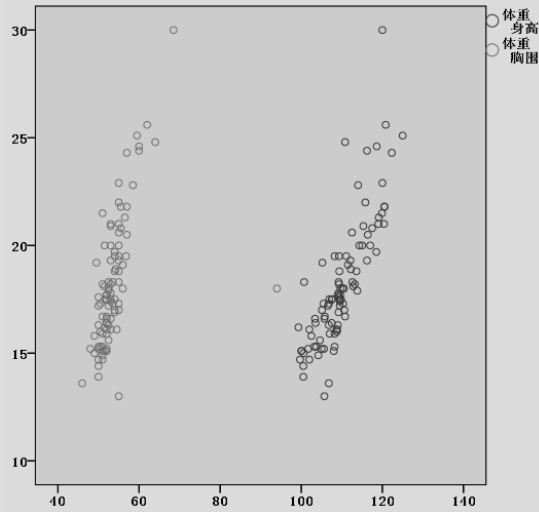


图 13-48 例 13-9 重叠散点图

从图中可以看出，体重与身高基本呈线性关系，随着身高的增加，体重也在增加；体重与胸围也一样呈线性关系。

13.4.11 直方图

直方图（Histogram）是用一组无间隔的直条图来表现频数分布特征的统计图，直方图的每一条形高度分别代表相应组别的频率。它是 SPSS 中一种很常见的图形，绘制也十分简单，以例 13-9 的数据为例，绘制儿童身高的直方图，具体步骤如下。

第 1 步 打开数据文件 data13-8.sav。

第 2 步 打开定义直方图的对话框。

选择“图形→旧对话框→直方图”，弹出如图 13-49 所示的“直方图”对话框。



图 13-49 “直方图”对话框

第 3 步 确定绘制直方图的变量。

从左边的变量列表框中选择变量“身高”，移入“变量”框中，同时再选中“显示正态曲线”复选框。

第 4 步 运行结果及说明。

单击如图 13-49 所示对话框中的“确定”按钮，完成直方图的绘制，得到图 13-50 的结果图。

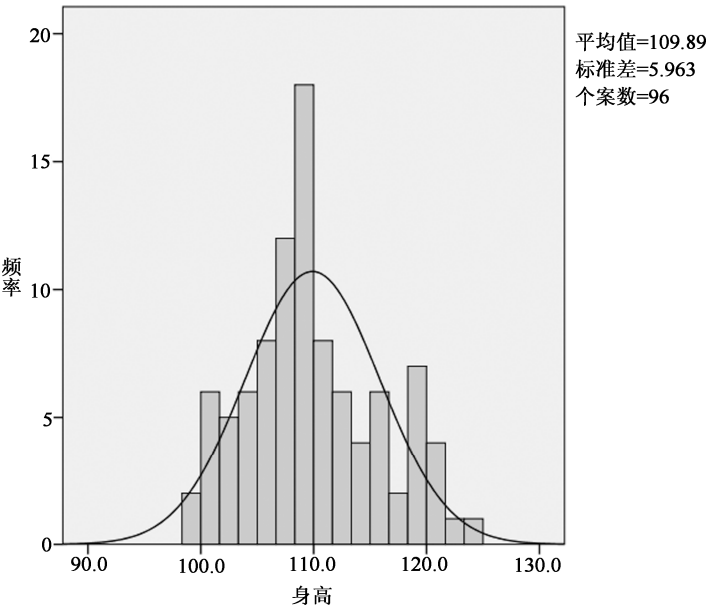


图 13-50 儿童身高直方图

如果数据样本足够大，优秀的统计工作者从图形上就可以判断变量满足的大致分布类型。

13.5 图表的编辑

图表生成后，还可以在结果浏览窗口里双击图表，启动图表编辑窗口，对图表进一步编辑和探索。

13.5.1 图表编辑器布局

图表编辑器打开后如图 13-51 所示，该窗口由菜单栏、工具栏和编辑区组成，通常图表编辑器打开后“属性”窗口同时打开，如图 13-52 所示，“属性”窗口根据用户当前的操作内容动态显示一些供用户修改和设置的内容，由多个选项卡组成。

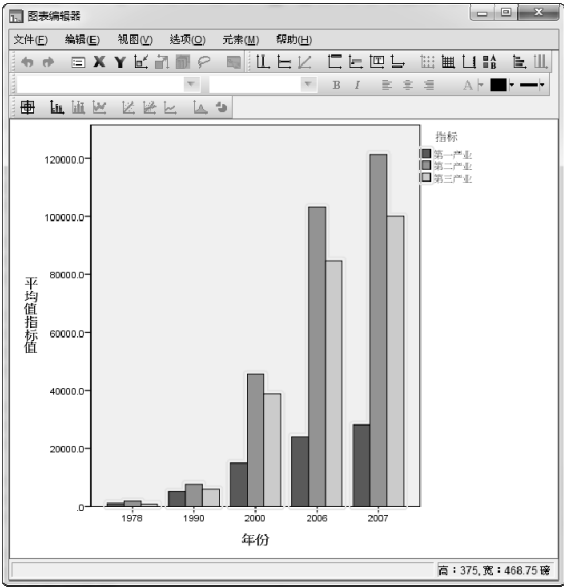


图 13-51 图表编辑器

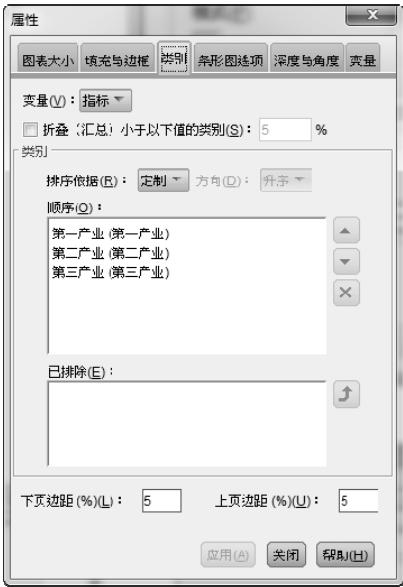


图 13-52 图形属性设置

图表编辑器各菜单主要作用如下。

- (1) 文件：保存和打开定义好的图形模板。图形模板是指保存用户自定义的图形大小、颜色等信息的文件。
- (2) 编辑：除了常见的撤销、重做、剪切、复制、粘贴外，主要包括图表选择、坐标轴选择、打开属性窗口、图表调整和 3D 旋转等。
- (3) 视图：主要进行工具栏、状态栏的打开和关闭切换。
- (4) 选项：主要对图表的辅助元素进行设置，如添加 X 轴、Y 轴参考线，添加标题、注释、文本框脚注，显示和隐藏网格线、派生轴，变换图表等。
- (5) 元素：主要添加一些附加图形元素，如数据标签、误差条图、拟合线等。
- (6) 帮助：SPSS 帮助菜单。

文件、视图和帮助菜单都比较简单和常见，图表编辑器的核心功能主要由编辑、选项和元素菜单完成，因此，SPSS 也把这些菜单放在相应的工具栏上，方便用户快捷操作。

13.5.2 图表编辑基本方法

图表编辑主要由以下 3 种方法结合来完成。

1. 菜单

图表编辑器中可以执行的很多操作都是用菜单完成的，特别是将某项添加到图表的时候。例如，使用菜单将拟合线添加到散点图。将某项添加到图表后，经常需要使用“属性”窗口来指定对应添加项的选项。

2. 工具栏

工具栏提供菜单和“属性”窗口中某些功能的快捷方式。例如，可以不用“属性”窗口的“文本”选项卡，而使用“编辑”工具栏来更改文本的字体和样式。

3. “属性”窗口

在“属性”窗口可以找到对应图表及其图表元素的选项。要打开“属性”窗口可以双击图表元素，或选择图表元素，然后在菜单中选择“编辑→属性”。此外，“属性”窗口在将某项添加到图表时自动出现。

使用“属性”窗口的选项卡可以设置选项，并对图表做出其他更改。在“属性”窗口显示的选项卡取决于当前的选择，如果选择了多个元素，则只能查看和更改所有元素共有的设置。

13.5.3 图表基本设定

1. 选择、移动图表元素和调整图表元素的大小

修改图表中的任何元素之前必须先选择这些元素，只有在选择了某个元素后才可以看见“属性”窗口中该元素的所有可用选项。选择图表元素主要通过鼠标单击来完成；有时候需要多次单击鼠标，逐次缩小选择范围来完成；还可以通过菜单来完成，如选择 X 轴、 Y 轴。除此之外，还可以用 Tab 键进行轮换选择，Ctrl 键+鼠标进行多个元素选择等。

选中图表元素后，鼠标移到选中元素的边缘上方，如果出现十字箭头，则可以按住鼠标将元素拖到新位置，也可以用键盘的方向键进行移动。

将光标放在元素边框用于调整大小的控柄上，出现一个双头的箭头时，单击并拖动调整大小的控柄直到元素的大小合适为止。还可以使用“属性”窗口的“图表大小”选项卡调整外框大小。

2. 更改图表的外观

更改图表的外观包括很多繁杂的项目：文本的内容、大小、字体、颜色和布局；填充和边框样式，如外框、数据框、图注框、文本框和条形图的线宽和边框样式；标记样式，如数据标记（点）的形状及其大小、颜色、边框宽度；以及各种线、条、饼等图形元素的样式，轴标题、刻度标记和刻度标记标签，网格线、数字、日期格式等。这些基本都是通过“属性”窗口相应的选项卡进行设置的，操作非常简单。

3. 在图表中添加和更改文本

图表中添加文本包括文本框、标题、脚注、注释、数据标签等，前 4 项主要通过“选项”菜单和工具栏完成，其菜单项和工具栏按钮都相邻放着，基本上和文本框操作一样。当添加这 4 种对象时，光标会在默认文本中闪烁，此时，就可以直接输入编辑文本。如果要输入多行文本，可

用 shift+回车键换行。数据标签由“元素”菜单下的“数据标签模式”进行添加和取消。选中上述对象，可以在“属性”窗口中对文本及边框的外观进行设置。

13.5.4 图表高级设定

1. 探索图表中的数据

可使用图表编辑器探索图表中的数据，深入发掘数据的分布规律及趋势等。可以添加内插线、拟合线和参考线，还可以将对数函数添加到轴来转换刻度轴等，这些操作主要通过“元素”菜单及工具栏完成，这些菜单项及功能按钮也是紧邻在一起的。对于 3D 图表，可以朝各个方向进行任意旋转或对其进行缩放。甚至可以将图表从一种类型转换到另一种类型。例如，可以将条形图转换为饼图，或将散点图转换为散点矩阵。

2. 使用图表模板

图表模板可以保存一个图表的设置，并将这些设置应用于其他图表。使用图表模板可自动创建图表的定制外观（例如共用的颜色方案）、定制设置（例如刻度轴范围）、定制选项（例如拟合线）等。

保存图表模板后，可以在图表编辑器中手动应用，或者在创建图表时指定模板。也可以指定一个默认的模板，该模板的设置将自动应用于所有图表。这些操作都在“文件”菜单中完成。

图表的编辑操作项目纷繁，但操作比较简单，限于篇幅，此处不再一一列举和示例。

13.6 思考与练习

1. 表 13.10 是各地区人口数及人口自然变动情况，根据表中的数据在 SPSS 中绘制如下图形。（数据来源：中国人口和就业统计年鉴 2007；参见数据文件：data13-9.sav。）

表 13.10 各地区人口数及人口自然变动情况

地区	出生率	死亡率	自然增长率	年末人口数（万人）
北京	6.26	4.97	1.29	1581
天津	7.67	6.07	1.60	1075
河北	12.82	6.59	6.23	6898
山西	11.48	5.73	5.75	3375
内蒙古	9.87	5.91	3.96	2397

- （1）各地区人口的出生率、死亡率对比条形图；
- （2）各地区人口自然增长率的折线图；
- （3）各地区年末人口数的饼图。

2. 表 13.11 是成都地区 2005 年和 2006 年生猪出栏量的数据，试绘制：

- （1）2005 年和 2006 年生猪出栏数据箱图，要求两年数据在同一图中显示；
- （2）两年数据散点图。

表 13.11 成都各地区生猪出栏量

地区	锦江区	青羊区	金牛区	成华区	龙泉驿区	青白江区	新都区	温江区
2005 年	1.87	1.59	5.04	6.29	41.43	32.33	52.09	34.9
2006 年	1.86	0.26	4.7	4.56	43.13	33.29	54.01	33.87

3. 表 13.12 是北京各区柳树健康测试的结果表的部分数据, 根据这份测试数据, 绘制如图 13-53 所示的饼图, 直观展示各区柳树的健康情况。(参见数据文件: data13-10.sav。)

表 13.12 柳树健康测试情况表

树木编号	测定地点	健康等级	树木编号	测定地点	健康等级
1	海淀区	一般健康	14	朝阳区	中度危害
2	海淀区	轻度危害	15	丰台区	非常健康
3	海淀区	轻度危害	16	丰台区	一般健康
4	海淀区	轻度危害	17	丰台区	轻度危害
5	海淀区	一般健康	18	通州区	非常健康
34	通州区	非常健康	19	通州区	非常健康
35	通州区	非常健康	20	丰台区	非常健康
8	海淀区	非常健康	21	丰台区	一般健康
9	海淀区	非常健康	22	顺义区	非常健康
10	海淀区	非常健康	23	顺义区	一般健康
11	朝阳区	非常健康	24	丰台区	非常健康
12	朝阳区	非常健康	25	丰台区	非常健康
13	朝阳区	轻度危害	26	通州区	非常健康

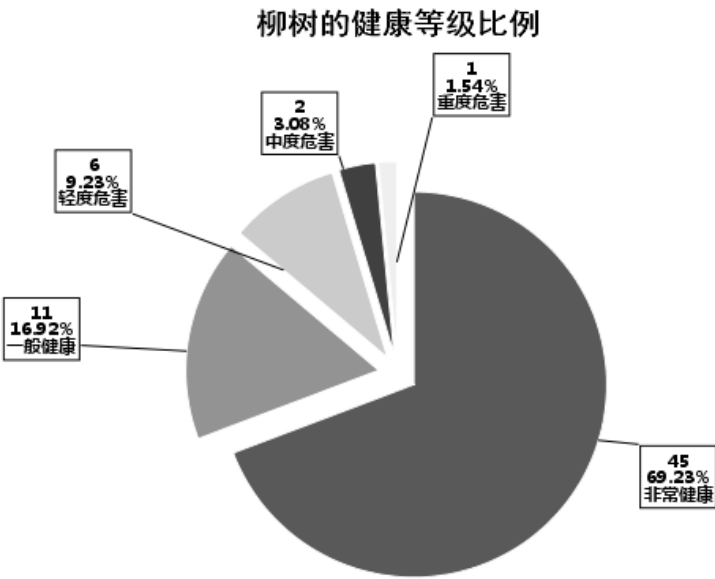


图 13-53 柳树健康等级比例图

参考文献

- [1] 贾俊平, 何晓群, 金勇进. 统计学(第六版). 北京: 中国人民大学出版社, 2015.
- [2] 周玉敏, 邓维斌. SPSS 16.0 与统计数据分析. 成都: 西南财经大学出版社, 2009.
- [3] 邓维斌, 唐兴艳, 周玉敏, 胡大权. SPSS 19(中文版) 统计分析实用教程. 北京: 电子工业出版社, 2012.
- [4] 王雪华, 张鹏, 张承伟. 管理统计学: 基于 SPSS 软件应用. 北京: 电子工业出版社, 2011.
- [5] 赖国毅, 陈超. SPSS 17.0 中文版常用功能与应用实例精讲. 北京: 电子工业出版社, 2010.
- [6] 卢纹岱, 朱红兵. SPSS 统计分析(第 5 版). 北京: 电子工业出版社, 2015.
- [7] 薛薇. SPSS 统计分析方法及应用(第 3 版). 北京: 电子工业出版社, 2013.
- [8] 郝黎仁, 樊元, 郝哲欧. SPSS 实用统计分析. 北京: 中国水利水电出版社, 2003.
- [9] 罗应婷, 杨钰娟. SPSS 统计分析从基础到实践. 北京: 电子工业出版社, 2007.
- [10] 宇传华. SPSS 与统计分析(第 2 版). 北京: 电子工业出版社, 2014.
- [11] 吕振通, 张凌云. SPSS 统计分析与应用. 北京: 机械工业出版社, 2010.
- [12] 吴明隆. 问卷统计分析实务——SPSS 操作与应用. 重庆: 重庆大学出版社, 2010.
- [13] 夏怡凡. SPSS 统计分析精要与实例详解. 北京: 电子工业出版社, 2010.
- [14] 李昕. SPSS 22.0 统计分析从入门到精通. 北京: 电子工业出版社, 2015.
- [15] 杨世莹, 高健. SPSS 22 统计分析案例教程. 北京: 中国水利水电出版社, 2016.
- [16] 冯岩松. SPSS 22.0 统计分析应用教程. 北京: 清华大学出版社, 2015.
- [17] 杨维忠, 张甜, 刘荣. SPSS 统计分析 with 行业应用案例详解. 北京: 清华大学出版社, 2015.
- [18] 李志辉, 罗平. SPSS 常用统计分析教程(SPSS 22.0 中英文版)(第 4 版). 北京: 电子工业出版社, 2015.
- [19] 陈胜可, 刘荣. SPSS 统计分析从入门到精通. 北京: 清华大学出版社, 2015.

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：(010) 88254396；(010) 88258888

传 真：(010) 88254397

E-mail: dbqq@phei.com.cn

通信地址：北京市海淀区万寿路 173 信箱

电子工业出版社总编办公室

邮 编：100036